

A COMPANION TO FREE WILL

EDITED BY
Joseph Campbell, Kristin M. Mickelson,
and V. Alan White



WILEY Blackwell

A Companion to Free Will

Blackwell Companions to Philosophy

This outstanding student reference series offers a comprehensive and authoritative survey of philosophy as a whole. Written by today's leading philosophers, each volume provides lucid and engaging coverage of the key figures, terms, topics, and problems of the field. Taken together, the volumes provide the ideal basis for course use, representing an unparalleled work of reference for students and specialists alike. For the full list of series titles, please visit wiley.com.

1. The Blackwell Companion to Philosophy, Second Edition
Edited by Nicholas Bunnin and Eric Tsui-James
2. A Companion to Ethics
Edited by Peter Singer
3. A Companion to Aesthetics, Second Edition
Edited by Stephen Davies, Kathleen Marie Higgins, Robert Hopkins, Robert Stecker, and David E. Cooper
4. A Companion to Epistemology, Second Edition
Edited by Jonathan Dancy, Ernest Sosa, and Matthias Steup
5. A Companion to Contemporary Political Philosophy (two-volume set), Second Edition
Edited by Robert E. Goodin and Philip Pettit
6. A Companion to Philosophy of Mind
Edited by Samuel Guttenplan
7. A Companion to Metaphysics, Second Edition
Edited by Jaegwon Kim, Ernest Sosa, and Gary S. Rosenkrantz
8. A Companion to Philosophy of Law and Legal Theory, Second Edition
Edited by Dennis Patterson
9. A Companion to Philosophy of Religion, Second Edition
Edited by Charles Taliaferro, Paul Draper, and Philip L. Quinn
10. A Companion to the Philosophy of Language, Second Edition (two-volume set)
Edited by Bob Hale and Crispin Wright
11. A Companion to World Philosophies
Edited by Eliot Deutsch and Ron Bontekoe
12. A Companion to Continental Philosophy
Edited by Simon Critchley and William Schroeder
13. A Companion to Feminist Philosophy
Edited by Alison M. Jaggar and Iris Marion Young
14. A Companion to Cognitive Science
Edited by William Bechtel and George Graham
15. A Companion to Bioethics, Second Edition
Edited by Helga Kuhse and Peter Singer
16. A Companion to the Philosophers
Edited by Robert L. Arrington
17. A Companion to Business Ethics
Edited by Robert E. Frederick
18. A Companion to the Philosophy of Science
Edited by W. H. Newton-Smith
19. A Companion to Environmental Philosophy
Edited by Dale Jamieson
20. A Companion to Analytic Philosophy
Edited by A. P. Martinich and David Sosa
21. A Companion to Genetics
Edited by Justine Burley and John Harris
22. A Companion to Philosophical Logic
Edited by Dale Jacquette
23. A Companion to Early Modern Philosophy
Edited by Steven Nadler
24. A Companion to Philosophy in the Middle Ages
Edited by Jorge J. E. Gracia and Timothy B. Noone
25. A Companion to African-American Philosophy
Edited by Tommy L. Lott and John P. Pittman
26. A Companion to Applied Ethics
Edited by R. G. Frey and Christopher Heath Wellman
27. A Companion to the Philosophy of Education
Edited by Randall Curren
28. A Companion to African Philosophy
Edited by Kwasi Wiredu
30. A Companion to Rationalism
Edited by Alan Nelson
31. A Companion to Pragmatism
Edited by John R. Shook and Joseph Margolis
32. A Companion to Ancient Philosophy
Edited by Mary Louise Gill and Pierre Pellegrin
35. A Companion to Phenomenology and Existentialism
Edited by Hubert L. Dreyfus and Mark A. Wrathall
39. A Companion to the Philosophy of Biology
Edited by Sahotra Sarkar and Anya Plutynski
41. A Companion to the Philosophy of History and Historiography
Edited by Aviezer Tucker
43. A Companion to the Philosophy of Technology
Edited by Jan-Kyrre Berg Olsen, Stig Andur Pedersen, and Vincent E. Hendricks
44. A Companion to Latin American Philosophy
Edited by Susana Nuccetelli, Ofelia Schutte, and Otávio Bueno
45. A Companion to the Philosophy of Literature
Edited by Garry L. Hagberg and Walter Jost
46. A Companion to the Philosophy of Action
Edited by Timothy O'Connor and Constantine Sandis
47. A Companion to Relativism
Edited by Steven D. Hales
50. A Companion to Buddhist Philosophy
Edited by Steven M. Emmanuel
52. A Companion to the Philosophy of Time
Edited by Heather Dyke and Adrian Bardon
64. A Companion to Experimental Philosophy
Edited by Justin Sytsma and Wesley Buckwalter
65. A Companion to Applied Philosophy
Edited by Kasper Lippert-Rasmussen, Kimberley Brownlee, and David Coady
70. A Companion to Nineteenth-Century Philosophy
Edited by John Shand
71. A Companion Atheism and Philosophy
Edited by Graham Oppy
78. A Companion to Free Will
Edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White

A Companion to Free Will

Edited by
Joseph Campbell,
Kristin M. Mickelson, and
V. Alan White

WILEY Blackwell

This edition first published 2023
© 2023 John Wiley & Sons, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right Joseph Campbell, Kristin M. Mickelson, and V. Alan White of to be identified as the authors of the editorial material in this work has been asserted in accordance with law.

Registered Office

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office

111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

The contents of this work are intended to further general scientific research, understanding, and discussion only and are not intended and should not be relied upon as recommending or promoting scientific method, diagnosis, or treatment by physicians for any particular patient. In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of medicines, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each medicine, equipment, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Campbell, Joseph - editor.

Title: A companion to free will / edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White.

Description: Hoboken, NJ, USA : Wiley Blackwell, [2023] | Series: Blackwell companions to philosophy |

Includes bibliographical references and index.

Identifiers: LCCN 2022040331 (print) | LCCN 2022040332 (ebook) | ISBN 9781119210139 (hardback) |

ISBN 9781119210153 (epdf) | ISBN 9781119210160 (epub) | ISBN 9781119210177 (ebook)

Subjects: LCSH: Free will and determinism.

Classification: LCC BJ1461 .C54 2023 (print) | LCC BJ1461 (ebook) | DDC 297.2/27--dc23/eng/20221103

LC record available at <https://lcn.loc.gov/2022040331>

LC ebook record available at <https://lcn.loc.gov/2022040332>

Cover image: © Rosenwald Collection/Wikimedia

Cover design by Wiley

Set in 10/12.5pt Photina by Integra Software Services Pvt. Ltd, Pondicherry, India

For Ella Mae (Moore) White
who modeled φίλος and σοφῶς in life

Contents

<i>List of Contributors</i>	x
<i>Acknowledgments</i>	xvi
1 Introduction, <i>Wiley Companion to Free Will</i> <i>Kristin M. Mickelson, Joseph Campbell, and V. Alan White</i>	1
Part I Preliminaries	21
2 Logical and Theological Fatalism <i>Alicia Finch</i>	23
3 Causal Determinism <i>V. Alan White</i>	39
4 (In)compatibilism <i>Kristin M. Mickelson</i>	58
5 Agent Causation <i>Leigh C. Vicens</i>	84
6 Obligation and Moral Responsibility <i>Ishtiyaque Haji</i>	95
7 Perfect Freedom <i>Marilyn Mccord Adams</i>	108
Part II Compatibility Problems	123
8 The Consequence Argument and the <i>Mind</i> Argument <i>Joseph Campbell and Kenji Lota</i>	125
9 Manipulation and Direct Arguments <i>Justin A. Capes</i>	144
10 Freedom and Time Travel <i>Ryan Wasserman</i>	157
11 Divine Freedom <i>Brian Leftow</i>	169
12 Denialism <i>Saul Smilansky</i>	184

CONTENTS

13	Revisionism <i>Manuel R. Vargas</i>	204
Part III The Science of Free Will		221
14	How the Laws Constrain: Causation, Counterfactuals, and Free Will <i>Kadri Vihvelin</i>	223
15	Free Will and Implicit Attitudes <i>Neil Levy and Jessica Wright</i>	241
16	The Role of Consciousness in Free Action <i>Philip Woodward</i>	256
17	Neuroscience <i>R.R. Waller</i>	278
18	A Defense of Natural Compatibilism <i>Florian Cova</i>	294
19	Libertarianism <i>Mark Balaguer</i>	314
Part IV Moral Responsibility		335
20	Children and Moral Responsibility <i>Meghan Griffith</i>	337
21	The Epistemic Condition of Moral Responsibility <i>Philip Robichaud</i>	355
22	Forgiveness and the Emotions <i>Laura W. Ekstrom</i>	369
23	Free Will and Moral Luck <i>Robert J. Hartman</i>	378
24	Basic Desert and the Appropriateness of Blame <i>Kelly McCormick</i>	393
25	Criminal Responsibility <i>Ken Levy</i>	406
Part V The Future		415
26	The Experience of Free Agency <i>Oisín Deery and Eddy Nahmias</i>	417
27	The Future of the Causal Quest <i>Hannah Tierney</i>	434
28	Free Will and Reference <i>Shaun Nichols</i>	451

29	Meaning in Life and Free Will Skepticism <i>Derk Pereboom</i>	464
30	Free Will: Looking Ahead <i>Alfred R. Mele</i>	477
31	Epilogue: Free Will Zombies <i>V. Alan White</i>	491
	Index	496

List of Contributors

Mark Balaguer is professor of philosophy at California State University, Los Angeles. He is the author of four books – *Platonism and Anti-Platonism in Mathematics* (Oxford University Press, 1998), *Free Will as an Open Scientific Problem* (MIT Press, 2010), *Free Will* (MIT Press, 2014), and *Metaphysics, Sophistry, and Illusion: Toward a Widespread Non-Factualism* (Oxford University Press, 2021). He has also published numerous articles on a wide range of philosophical topics in journals such as *Mind*, *Nous*, and *Philosophy and Phenomenological Research*.

Joseph Campbell is professor of philosophy in the School of Politics, Philosophy, and Public Affairs at Washington State University. He is co-founder of the Inland Northwest Philosophy Conference and has helped organize scores of philosophy conferences and public events. Professor Campbell has edited nine books as well as numerous papers for the *Journal of Ethics*, *The Internet Encyclopedia of Philosophy*, and *Philosophical Studies*. He is the recipient of the Marian E. Smith Faculty Achievement Award and the Honors Thesis Advisor Award.

Justin Capes is assistant professor of philosophy in the School of Humanities and Science at Flagler College and an Associate Editor for the *Journal of Ethics*. He writes on topics in moral philosophy and the philosophy of action, especially those having to do with free will and moral responsibility. His published work on these issues has appeared in journals such as *Philosophy and Phenomenological Research*, *Philosophical Studies*, *American Philosophical Quarterly*, *Pacific Philosophical Quarterly*, *Erkenntnis*, *Social Philosophy & Policy*, among others. He is currently in the process of finishing a book defending the much-debated Principle of Alternative Possibilities.

Florian Cova is a visiting assistant professor at University Geneva, Switzerland. As an experimental philosopher, his research is at the intersection between philosophy and cognitive science and explores how we think about a multitude of philosophical issues: aesthetics, free will, intentional action, or morality. His current work investigates people's conceptions of the meaning of life and the role emotions play in our search for a meaningful life. He is the coeditor of *Advances in Experimental Philosophy of Aesthetics* and has published in journals including *Consciousness & Cognition*, *Mind and Language*, *Personality and Social Psychology Bulletin*, *Philosophical Psychology*, *Philosophical Studies* and the infamous *Asian Journal of Medicine and Health*.

Oisín Deery is ARC DECRA fellow and lecturer in the Department of Philosophy at Macquarie University, Sydney, and assistant professor in the Department of Philosophy, York University, Toronto. His research is primarily in the philosophy of mind and action. He is currently working on issues in artificial intelligence, including ethical issues related to agency, for a three-year

research project funded by the Australian Research Council (ARC). His monograph, *Naturally Free Action*, appeared in 2021 with Oxford University Press. In 2013, he published a coedited volume with Oxford University Press, *The Philosophy of Free Will: Essential Readings from the Contemporary Debates*. His articles have appeared in *Philosophical Studies*, *Australasian Journal of Philosophy*, *Synthese*, and *Philosophical Psychology*.

Laura W. Ekstrom is Francis S. Haserot Chancellor professor of philosophy at William & Mary. She is a graduate of Stanford University (B.A.) and the University of Arizona (PhD). Her books include *God, Suffering, and the Value of Free Will* (Oxford University Press, 2021) and *Free Will: A Philosophical Study* (Westview Press, 2000). Her articles on autonomy, moral responsibility, causation, chance, free will, compassion, and suffering have been published in academic journals including *Philosophy and Phenomenological Research*, *Synthese*, *The Journal of Medicine and Philosophy*, *American Philosophical Quarterly*, *Philosophical Studies*, *Midwest Studies in Philosophy*, and *Australasian Journal of Philosophy*, as well as in edited collections published by Blackwell, Routledge, Oxford University Press, and Cambridge University Press.

Alicia Finch is associate professor of philosophy at Northern Illinois University. Her research has focused on the metaphysics of free will and moral responsibility, and she has taught courses in metaphysics, meta metaphysics, philosophy of religion, moral psychology, ancient philosophy, feminism, and the philosophy of race. Prior to joining NIU's faculty, she was a post-doctoral research fellow at the Notre Dame Center for Philosophy of Religion and assistant professor of philosophy at Saint Louis University. She received her B.A. in Philosophy and Political Science from the University of Missouri at Columbia, and her M.A. and PhD in Philosophy from the University of Notre Dame.

Meghan Griffith is professor of philosophy at Davidson College. She is interested in moral responsibility, free will, the metaphysics of agency, and other topics related to human action. She has published a number of journal articles and book chapters on these topics. She is the author of *Free Will: The Basics*, 2nd Edition (Routledge, 2022), and a co-editor of the *Routledge Companion to Free Will* (Routledge, 2017). Recent research focuses on the role our narrative capacities play in the development of morally responsible agency.

Ishtiyaque Haji is professor of philosophy at the University of Calgary. He has research interests in ethical theory, philosophy of action, metaphysics, and philosophical psychology. He is the author of *Moral Appraisability* (1998), *Deontic Morality and Control* (2002), (with Stefaan Cuypers) *Moral Responsibility, Authenticity, and Education* (2008), *Freedom and Value* (2009), *Incompatibilism's Allure* (2009), *Reason's Debt to Freedom* (2012), *Luck's Mischief* (2016), *The Obligation Dilemma* (2019), and *Obligation and Responsibility* (2023).

Robert J. Hartman is an assistant professor of philosophy at Ohio Northern University. He works mainly on agency and responsibility, character and virtue, and philosophy of religion. Currently, he is writing a monograph titled *Character and Free Will*. He is the author of *In Defense of Moral Luck: Why Luck Often Affects Praiseworthiness and Blameworthiness* (Routledge 2017) and co-editor of *The Routledge Handbook of the Philosophy and Psychology of Luck* (Routledge 2019). His work also appears in journals such as *Australasian Journal of Philosophy*, *Erkenntnis*, *Faith and Philosophy*, *Journal of the American Philosophical Association*, *Philosophical Studies*, and *Thought*.

LIST OF CONTRIBUTORS

Brian Leftow is William P Alston professor of the philosophy of religion and co-director of the Rutgers Center for Philosophy of Religion, Department of Philosophy, Rutgers University. He is also an Emeritus Fellow of Oriel College Oxford. He is the author of *Anselm's Argument* (OUP, 2022), *God and Necessity* (OUP 2012), *Time and Eternity* (Cornell, 1991), and well over 100 articles in philosophy of religion, metaphysics, and medieval philosophy. For 16 years he was Nolloth Professor of the Philosophy of the Christian Religion, Oxford University, and a Fellow of Oriel.

Neil Levy is professor of philosophy at Macquarie University, Sydney, and a Senior Research Fellow at the Uehiro Centre for Practical Ethics, University of Oxford. He is a wide-ranging philosopher, who has published books and articles on free will, social epistemology, applied ethics, philosophy of mind, and other topics. His major work on free will is *Hard Luck: How Luck Undermines Free Will and Moral Responsibility* (Oxford University Press, 2011).

Ken M. Levy is the Holt B. Harrison professor of law at LSU Law School in Baton Rouge, Louisiana. He teaches courses in criminal law and has published articles in several areas, including free will and responsibility, criminal theory, and constitutional law. He recently published a book entitled *Free Will, Responsibility, and Crime: An Introduction* (Routledge, 2020).

Kenji Lota is a graduate student in the Department of Philosophy at the University of Miami. Before coming to the University of Miami, he completed his master's degree in philosophy at Concordia University in Montreal. His research interests are in epistemology, philosophy of language, and philosophy of action. His current work explores different ways of understanding interrogative attitudes, the norms that govern those attitudes, and how it relates to free will and action. He has coauthored papers that will appear in *Erkenntnis* and *Ergo*.

Marilyn McCord Adams (1943–2017) spent the longest part of her career as a philosophy professor at the University of California–Los Angeles (1972–1993). She moved from there to the Divinity School at Yale, as professor of the history of Christian doctrine in the medieval and early modern periods (1993–2003). Crossing the ocean, she became the first woman to hold the post of Regius Professor of Divinity at Oxford University – the first woman to hold that historic post. Her major books were *William Ockham*, in two volumes (1987), *Horrendous Evils and the Goodness of God* (1999), *Christ and Horrors: The Coherence of Christology* (2006), and *Some Later Medieval Theories of the Eucharist* (2010).

Kelly McCormick is an associate professor of philosophy at Texas Christian University. Her research primarily concerns free will and moral responsibility, with special emphasis on the permissibility of blame and the variety of methodological issues that give rise to disagreement between eliminativists and preservationists. Her first monograph on these issues, *The Problem of Blame: Making Sense of Moral Anger*, will be published by Cambridge University Press in 2022. Her publications also appear in journals including *Philosophical Studies*, *Social Philosophy and Policy*, *Criminal Justice Ethics*, *The Journal of Value Inquiry*, and *The Journal of Ethics and Social Philosophy*.

Alfred R. Mele is the William H. and Lucyle T. Werkmeister Professor of Philosophy at Florida State University. He is the author of thirteen books and over 250 articles and editor of seven books. He is past director of two multi-million-dollar, interdisciplinary projects: the Big Questions in Free Will project (2010–13) and the Philosophy and Science of Self-Control project (2014–17). His latest book is *Free Will: An Opinionated Guide* (Oxford University Press, 2022). Free will is one of his favorite topics.

Kristin M. Mickelson earned her Ph.D. from the University of Colorado at Boulder and spent two years as a postdoctoral researcher at the University of Gothenburg (Sweden) in the Lund-Gothenburg Responsibility Project. She is now an independent researcher who works on the metaphysics of free will (broadly construed), the logic of explanation, and moral luck. She is particularly interested in redressing anomalies and stalemates that emerged in the free-will debate during the degeneration of the classical analytic paradigm (CAP). Towards that end, she has developed an alternative (non-classical) way of framing the problem of free will – she calls it “The Paradox Paradigm” – which opens up new avenues of research by exposing the dialectical connections between famous paradoxes of control, including the paradox of (in) determinism, the paradox of self-creation, and the paradox of moral luck. Her published work appears in venues such as *Australasian Journal of Philosophy*, *Midwest Studies in Philosophy*, *Philosophia*, *Canadian Journal of Philosophy*, and *Social Philosophy & Policy*.

Eddy Nahmias is professor and Chair of philosophy at Georgia State University. His research is devoted to the study of human agency: what it is, how it is possible, and how it accords with scientific accounts of human nature. His primary focus is the free-will debate, including potential threats to free will posed by the sciences of the mind, and conversely, what these sciences can illuminate how free will works in humans. He has published over 40 articles in these areas, and is co-editor of *The Natural Method: Essays on Mind, Self, and Ethics in Honor of Owen Flanagan* and *Moral Psychology: Historical and Contemporary Readings*.

Shaun Nichols is professor of philosophy at Cornell University. His research concerns the psychological underpinnings of philosophical thought. He is the author of *Sentimental Rules: On the Natural Foundations of Moral Judgment*, *Bound: Essays on Free Will and Moral Responsibility*, and *Rational Rules: Towards a Theory of Moral Learning*, and he has published over 100 articles in academic journals in philosophy and psychology.

Derk Pereboom is the Susan Linn Sage Professor in the department of philosophy at Cornell University. His areas of research include free will and moral responsibility, philosophy of mind, and early modern philosophy, especially Kant. He is the author of *Living without Free Will* (Cambridge 2001), *Consciousness and the Prospects of Physicalism* (Oxford 2011), *Free Will, Agency, and Meaning in Life* (Oxford 2014), and *Wrongdoing and the Moral Emotions* (Oxford 2021). He has published articles on free will and moral responsibility, consciousness and physicalism, nonreductive materialism, philosophy of religion, and on Kant’s metaphysics and epistemology.

Robyn Repko Waller is a lecturer in Philosophy at the University of Sussex. Her primary research interests lie in philosophy of mind and philosophy of cognitive science. She has published papers on agency, free will, causation, moral responsibility, neuroethics, and biomedical ethics in journals such as the *Monist*, *Philosophical Psychology*, *Grazer Philosophische Studien*, *Synthese*, and the *Cambridge Quarterly of Healthcare Ethics*, as well as forthcoming chapters in academic press anthologies.

Philip Robichaud is associate professor of philosophy at the Vrije Universiteit Amsterdam. His areas of research specialization are moral responsibility, value theory, and the ethics of behavioral interventions, such as nudges. He coedited with Jan Willem Wieland the edited volume *Responsibility: The Epistemic Condition*, and his journal publications have appeared in *Ethics*, *Philosophy and Phenomenological Research*, *The Monist*, and *Science and Engineering Ethics*.

LIST OF CONTRIBUTORS

Saul Smilansky (D.Phil., Oxford) is a professor at the Department of Philosophy, University of Haifa, Israel. He works primarily on normative and applied ethics, the free will problem, and meaning in life. Currently he is working on the idea of “Crazy Ethics” where matters seem true (or at least plausible) yet absurd. He is the author of *Free Will and Illusion* (Oxford University Press 2000), *10 Moral Paradoxes* (Blackwell 2007), and one hundred papers in journals and edited collections.

Hannah Tierney is an assistant professor in the philosophy department at the University of California, Davis. She has broad philosophical interests, and writes mainly on issues of free will, moral responsibility, and personal identity. Her publications appear in journals including *Analysis*, *Australasian Journal of Philosophy*, *Journal of Philosophy*, and *Philosophical Studies*.

Manuel Vargas is professor of philosophy at the University of California San Diego. He is the author of *Building Better Beings: A Theory of Moral Responsibility* and a coauthor of *Four Views on Free Will*. He works on questions at the intersection of agency and morality, and on topics in the history of Latin American philosophy.

Leigh Vicens is associate professor of philosophy at Augustana University in Sioux Falls, SD. Her research interests include metaphysical and ethical questions related to human freedom and moral responsibility, and related topics in philosophy of religion such as divine providence and the problem of evil. She is the coauthor of *God and Human Freedom* with Simon Kittle (Cambridge University Press, 2019) and coeditor of *Theological Determinism: New Perspectives* with Peter Furlong (Cambridge University Press, forthcoming). Her publications include articles in a number of journals, and she is currently Book Review Editor at *Faith and Philosophy*.

Kadri Vihvelin is professor of philosophy at the University of Southern California. She has a longstanding interest in puzzles about free will and freedom of action, and has defended controversial claims about both. About free will, she argues that the victim of Frankfurt’s counterfactual intervener retains free will despite his counterfactual shackles. About freedom of action, she argues that common sense is right and philosophical orthodoxy wrong: Lewis’s time traveling Tim really cannot kill Grandfather. Her book, *Causes, Laws, and Free Will: Why Determinism Doesn’t Matter*, was published in 2013 by Oxford University Press. Her work on topics at the intersection of metaphysics and ethics, including causation, dispositions, and the doing/allowing distinction, has appeared in *The Journal of Philosophy*, *Australasian Journal of Philosophy*, and *Canadian Journal of Philosophy*, among others.

Ryan Wasserman is professor of philosophy at Western Washington University. His main interests are in metaphysics, ethics, and the philosophy of language. He has published papers in many journals, including *Mind*, *Noûs*, *Philosophical Quarterly*, *Philosophical Studies*, and *Philosophy and Phenomenological Research*. He is also the author of *Paradoxes of Time Travel* (Oxford University Press, 2018) and the coeditor of *Metametaphysics* (Oxford University Press, 2008) and has written on various issues relating to free will.

V. Alan White is professor of philosophy Emeritus at the University of Wisconsin–Green Bay, Manitowoc Campus. He is the author of numerous articles in metaphysics as well as pedagogy, appearing in *Analysis*, *Erkenntnis*, *Philosophy*, *The Southern Journal of Philosophy*, *Process Studies*, and *Teaching Philosophy*, among many other journals and chapters in books. He also

for many years has written and performed dozens of philosophical parodies on his internet site *Philosophy Songs*, where he invites feedback for “encouragement or forgiveness.” He was the 1996 recipient of a Carnegie/CASE teaching award as Wisconsin Professor of the Year, and in 2009 of the University of Wisconsin Colleges Chancellor’s Award for Excellence, presented in Madison, Wisconsin.

Philip Woodward is assistant professor of Philosophy at Niagara University. His research is predominantly in the philosophy of mind and philosophical anthropology, focusing especially on the connection between consciousness and various other aspects of personhood, including intentionality, rationality, and agency. For academic year 2021–2022 he was awarded a residential research fellowship at the Henry Center for Theological Understanding. His publications appear in journals such as *Australasian Journal of Philosophy*, *Canadian Journal of Philosophy*, *Philosophical Psychology*, and *Phenomenology & Mind*.

Jessica Wright is an ethicist and policy analyst interested in bias, responsibility, and health policy. She completed her PhD thesis in philosophy, *Owning Implicit Attitudes*, at the University of Toronto. She is currently working on questions at the intersection of ethics and health policy, with a focus on health equity.

Acknowledgments

The editors thank the contributors for their chapters, assistance, and patience. For written comments on chapters, the editors thank Mark Balaguer, Gregg Caruso, Taylor Cyr, Chris Franklin, Robert Hartman, Richard Holton, Benjamin Matheson, Michael McKenna, Philip Swenson, Kevin Timpe, Neal Tognazzini, Jason Turner, Leigh Vicens, and especially Neil Levy.

Joe Campbell thanks The Hayek Fund of the Institute for Humane Studies for a spring 2022 grant, and the School of Politics, Philosophy, and Public Affairs at Washington State University for several Summer Research Grants to help support work on this volume.

Lastly, the editors thank Wiley-Blackwell and their staff, especially Durgadevi Shanmugasundaram, Charlie Hamlyn, and Will Croft.

Introduction, *Wiley Companion to Free Will*

KRISTIN M. MICKELSON, JOSEPH CAMPBELL,
AND V. ALAN WHITE

Brackets are used to refer to chapters in this volume (e.g. [10] refers to Chapter 10 of the volume).
A glossary of **bolded** terms is provided at the end of this chapter.

We wish this volume to be a sure companion to the study of free will, broadly construed to include action theory, moral and legal responsibility, and cohort studies feathering off into adjacent fields in the liberal arts and sciences. In addition to general coverage of the discipline, this volume attempts a more challenging and complementary accompaniment to many familiar narratives about free will. In order to map out some directions such accompaniment will take, in this introduction we anchor the thirty contributions to this volume in some common history from which they arise, and attempt to indicate where future work in free will and moral responsibility will—and has already begun to—depart from that history.

1 Preliminaries: Free Will and Determinism

The concept of free will is fraught with controversy, as readers of this volume likely know. Philosophers disagree about what free will is, whether we have it, what mitigates or destroys it, and what (if anything) it's good for. Indeed, philosophers even disagree about how to fix the referent of the term 'free will' for purposes of describing and exploring these disagreements (Nichols [28]). What one person considers a reasonably neutral working definition of 'free will' is often considered question-begging or otherwise misguided by another. Such disputes make it difficult to summarize the **problem of free will**, roughly the debate over the nature and existence of free will, in a clear and uncontentious way. In generic terms, however, the two basic solutions to the problem of free will are **free-willism**, the view that we (ordinary humans) have free will, and **free-will denialism**, i.e. the view that we do not have free will (Smilansky [12]).¹ As stated here, neither denialism nor free-willism constitutes a complete solution to the problem of free will; to be complete, a proposed solution must also tell us a convincing story about what free will *is* (what 'free will' *means*) and that, as it turns out, is a very difficult task indeed.

One historically popular way of approaching the problem of free will is to ask about the *relationship* between free will and determinism: “Does free will stand in relation *R* to determinism: *yes* or *no*?” This is just a template for a question, of course. To transform this template-question into a substantive question with a clear meaning, we need to flesh out the template’s free-will relatum, its determinism relatum, and give a precise value to relation *R*. There is, however, no uncontroversial way to do this. In addition to the standard difficulties raised by fixing the referent of ‘free will’, philosophers hold radically different views about what is – or should be – meant by the term ‘determinism’ (White [3], Vihvelin[14]; see also Beebe and Mele 2002, Shabo 2010), and they identify relations which are as substantively different as correlation and causation when characterising relation *R* (Mickelson [4]). In practical terms, then, it may be best to think of **the problem of determinism** as a loose collection of disagreements about how to best spell out and answer the template-question, and how (if at all) asking and answering such questions would help us to solve the problem of free will.

The term ‘determinism’ was ushered into the free-will literature in the 19th century, but the doctrine may be traced back to the Stoic’s naturalistic cause-and-effect theory of fate (Bobzien 1998, 2021), which may be contrasted with logical and theological varieties of fate which have also been of traditional interest vis-a-vis free will (Finch [2]). William James, in his influential “Dilemma of Determinism,” tells his audience that “no ambiguities hang about this word [determinism] or about its opposite, indeterminism” (James 1884). According to James, determinism “professes that those parts of the universe already laid down absolutely appoint and decree what the other parts shall be” such that the “future has no ambiguous possibilities hidden in its womb.” Indeterminism, says James, is true whenever “the parts have a certain amount of loose play on one another, so that the laying down of one of them does not necessarily determine what the others shall be” (James 1884). Put another way, **traditional determinism** (i.e. determinism as it was traditionally conceived within the free-will debate) is the doctrine that there is a causal or nomological necessity in nature which makes one unique future inevitable given what preceded it. **Traditional indeterminism** is the negation of traditional determinism; it is true if and only if it is false that one unique future is inevitable relative to any arbitrary moment in time (holding fixed the naturalistic factors which account for the evolution of the physical universe and the facts of the past, if any, relative to that time) (e.g. van Inwagen 1990, p. 277). Hereafter, we use ‘determinism’ as shorthand for traditional determinism and ‘indeterminism’ as shorthand for traditional indeterminism, unless stated otherwise.

Faced with the idea of determinism, many people, especially those working within the Christian tradition (Adams [7]), have argued that no one could exercise free will in a world at which this necessity-in-nature doctrine of predetermination is true; others—including the ancient Stoics—have disagreed. One popular way of tracking this age-old dispute has been to divide philosophers based on their answers to two questions: (1) “Is determinism true?” and (2) “Do we—ordinary humans—have free will?”. Those who answered “yes” to the first question were classified as *determinists*, and they were subdivided based on their preferred response to the second question. Determinists who answered “no” were classified as *hard* determinists, while those who answered “yes” were classified as *soft* determinists (James 1884). The determinists were contrasted primarily with *libertarians*, i.e. philosophers who answered “no” to the first question and “yes” to the second.² The term ‘hard’ in hard determinism indicates that some species of denialism is true. Despite their substantive differences, the soft determinists and libertarians agreed that denialism is false, which is to say that they agreed that free-willism is true.

In an innovative move, a group of philosophers working in the so-called *classical period* of the free-will debate, c. 1965–1985 (van Inwagen 2017), shifted the focal point away from the question of whether determinism is true to more theoretical questions which (according to their diagnosis) lay just under the surface of the pre-classical taxonomy of free-will views. One question was singled out as particularly important:

Is there a *conflict* or *tension* between the very notions of *free will* and *determinism*, such that if determinism were true, it would follow that determinism-related factors, i.e. the causal and/or nomological factors described by determinism, *preclude* free will (as the hard determinists and libertarians propose) or is there no such conflict (as the soft determinists believed)?

To raise the same question in slightly different terms, we could say—following the popular practice of using “luck” as shorthand for *factors beyond one’s control* (Hartman [23], Mickelson [4])—that these theorists were focused on the narrow question of whether or not *determinism-related causal luck* poses a distinct threat to free will. The challenge of answering the question of whether determinism (i.e. determinism-related causal factors beyond one’s control, determinism-related causal luck) precludes human free will is now widely known as **The Compatibility Problem**.

The Compatibility Problem was initially nested within the dominant research paradigm of the classical period: the **classical analytic paradigm (CAP)** (Mickelson [4]). Among CAP’s defining background assumptions, the assumption of **classical possibilism** is especially significant. Classical possibilism may be understood as the conjunction of two claims: (1) the **classical account of free will** is correct, i.e. free will is (some kind of) an ability to do otherwise, and (2) **anthropocentric possibilism**, the view that it is metaphysically possible for an ordinary human to exercise free will, is true (e.g. van Inwagen 1983, Clarke 2003, Vihvelin 2013, Mickelson [4]).

By assuming classical possibilism, CAP theorists (i.e. philosophers working within CAP) restricted the compatibility problem to the **classical compatibility problem**, roughly a debate about which possibilist interpretation of the ability to do otherwise is best. On the one hand, **classical incompatibilists** (e.g. Peter van Inwagen 1983) contend that a person exercises the ability to do otherwise (a.k.a. free will) when performing an action only if there is some kind of *indeterministic leeway* in the evolution of the physical world (see Smilansky [12], Balaguer [19]).³ Since all classical theorists accept classical possibilism, classical incompatibilism comes bundled with the endorsement of a broadly libertarian account of free will.⁴ As such, it is easy for the classical incompatibilist to explain why there is a deep conceptual *conflict* or an antagonistic *incompatibility* relation between the notions of free will and determinism. Since determinism states that there are naturalistic factors (i.e. determinism-related factors) which eliminate all indeterministic leeway in the evolution of the world, to say that determinism is true is to say that the world includes factors which eliminate the type of indeterministic leeway (whatever type that may be) that an exercise of free will requires.⁵ On the other hand, **classical compatibilists** (e.g. Keith Lehrer 1990) argue that indeterministic leeway is not required to exercise the ability to do otherwise. According to the classical compatibilist, the mere fact that determinism-related factors rule out all indeterministic leeway does not mean—*pace* the classical incompatibilists—that determinism-related factors rule out the ability to do otherwise (a.k.a. free will). However, classical compatibilists do not merely reject the classical incompatibilists’ “causal factors” *explanation* for the purported fact that normal humans cannot act freely when determinism; they are also committed, given their assumption of classical possibilism, to a classical version of **compossibilism**, the view that it is metaphysically possible

for an ordinary human to act freely in a world at which determinism is true (Mickelson [4]). As such, the classical compatibility problem may be summarized as the challenge of settling which of two views, classical (compossibilist) compatibilism or classical (broadly libertarian) incompatibilism, is correct.⁶

Challenges to CAP have given rise to other perspectives on the compatibility problem and to fundamentally different interpretations of the problem of determinism. Three challenges are worth noting, given their profound impact on the trajectory of the recent history of the free-will debate. Two of these challenges target the CAP assumption that classical possibilism is true, while the third challenges CAP's implicit practice of framing the problem of determinism as a narrow dispute about the relationship between free will and causal luck.

The first major strike against classical possibilism came in the form of Harry Frankfurt's influential criticisms of the classical (a.k.a. leeway) account of free will (Frankfurt 1969) specifically what are now known as "Frankfurt examples" (Haji [6]). By casting doubt on the classical conception of free will, Frankfurt examples motivated interest in non-classical accounts of free will, especially *sourcehood* accounts (Haji [6], Capes [9]). This, which is perhaps the most well-known critique of CAP's background assumptions, was not considered a fatal flaw in the CAP approach to the problem of determinism. Rather, it led philosophers to think that the term 'free will' should not be narrowly defined to mean "an ability to do otherwise" in generic statements of the compatibility problem. In such contexts, 'free will' should instead be defined in a way that opens dialectical space for a lively debate about which account of free will is correct.

This shift in the working definition of 'free will' led to the popularization of the **neo-classical compatibility problem**, which is (at least superficially) just like the classical compatibility problem except that the term 'free will' is used more broadly. The neo-classical use of 'free will' allows that the classical account of free will may be true, but it also allows that some non-classical account (e.g. a sourcehood account) may be correct. The two recognized solutions to the neo-classical compatibility problem are **neo-classical incompatibilism**, the view that is metaphysically impossible because determinism-related factors undermine free will (neo-classically defined) in worlds at which determinism is true, and **neo-classical compatibilism**, the view that determinism-related factors pose no threat to free will and it is metaphysically possible for an ordinary human to exercise free will (neo-classically defined) in a world at which determinism is true.⁷ Criticism of the classical definition of 'free will' also contributed to the centralization of *moral responsibility* in neo-classical and non-classical definitions of the term 'free will' (Haji [6], McCormick [24]), a point that we return to below (Section 4).⁸

As sourcehood accounts became mainstream, they helped to normalize the idea that, *pace* CAP theorists, anthropocentric possibilism may be false. While some source theorists, including Frankfurt, became neo-classical *source compatibilists* (arguing that it is possible for an ordinary human to satisfy the necessary source condition on free will even when determinism is true, from which it follows that determinism-related factors do not always undermine free will), other source theorists became neo-classical *source incompatibilists* (arguing that determinism-related factors preclude free will in virtue of keeping people from satisfying the source condition—as opposed to the classical ability-to-do-otherwise condition) on free will. Some of these neo-classical source incompatibilists, e.g. Derk Pereboom 2001, 2014, were also concerned about apparent threats to free will posed by *indeterministic* causal factors (i.e. indeterministic forms of causal luck). Such concerns led to the emergence of Pereboom's **hard source incompatibilism**, a species of **anthropocentric impossibilism** which claims that it is metaphysically impossible for an ordinary human (i.e. someone like us, as we are here on Earth) to exercise free will on the grounds that, whether determinism is true or false, some kind of causal luck (i.e. causal factors beyond one's control) ensures that no normal

human satisfies the necessary source condition on free will.⁹ Since hard source incompatibilism clearly speaks against both tenets of classical possibilism—rejecting both the classical account of free will and the assumption of anthropocentric possibilism—it is a decisively non-CAP position.

To be clear, Pereboom's hard source incompatibilism is not an example of full-blooded **impossibilism**, the unqualified view that it is impossible for anyone—even God (Adams [7]; Leftow [11])—to exercise free will. Pereboom is sympathetic to a broadly agent-causal (as opposed to event-causal) libertarian account of free will (e.g. Pereboom 2001, 2014; Vicens [5]) and this keeps him from endorsing unqualified impossibilism. Hard source incompatibilism is an influential view in part because it promises to provide a complete solution to the problem of free will: the “source” part tells us what free will amounts to and the “hard” part signals its endorsement of denialism. The hard source incompatibilist route to solving the problem of free will is attractive, in part, because it allows its proponents to adopt denialism without taking a stand on the truth-value of determinism.

The growing popularity of source accounts of free will has also raised the profile of philosophers who have been arguing for unqualified **impossibilism**. Among impossibilists (e.g. Galen Stawson 1986, Levy 2011, Mickelson 2019b), some argue for the radically anti-CAP view that the specific factors beyond our control which keep us from acting freely are not located in our environment (e.g. states in the remote past or the laws of nature) but are instead located *entirely* in facts about us. Drawing again on the language of “luck,” these source impossibilists contend that causal luck is irrelevant to free will. They claim, instead, that *constitutive luck*—roughly luck in the way that one is constituted, especially in regards to how one is mentally (at least in certain key respects), at the time of action—keeps people from acting freely, no matter what one's environment is like. This **constitutive-luck source impossibilism**, like its rival hard (source) incompatibilism, provides a route to denialism which does not require us to resolve tricky empirical questions about whether determinism is true or false.

Since constitutive-luck source impossibilism is in direct conflict with all three of the CAP tenets discussed above, it is a paradigmatically non-CAP position. It should not be surprising, then, that this view defies classification in CAP-based terms (e.g. Vihvelin 2008, McKenna and Pereboom 2016, p. 151, Mickelson [4]). Since these impossibilists reject the **compossibilist** component of classical/neo-classical compatibilism, they are not “compatibilists” in any traditional sense; but these impossibilists are not “incompatibilists” in the traditional sense either, for they also reject the explanatory tenet of classical/neo-classical incompatibilism which identifies determinism-related factors (i.e. causal luck) as relevant to free will. Just as there is a clear sense in which constitutive-luck source impossibilism is both an *anti-compatibilist* and *anti-incompatibilist* position, there is also a sense in which it is both a compatibilist and incompatibilist position: it is *incompatibilist* insofar as it entails the modest impossibilist tenet of traditional forms of incompatibilism, but *compatibilist* insofar as it denies that determinism-related factors pose a threat to free will (for further discussion, see Mickelson [4], 2015a, 2019b). Philosophers have yet to reach a consensus on whether—and, if so, how—to update CAP-based jargon so that it tracks non-CAP views.¹⁰

The chapters in this volume reflect a variety of classical, neo-classical, and non-classical perspectives on the problem of free will and the problem of determinism. While CAP remains a powerful and popular research framework, alternative approaches promise to raise new questions and inspire fruitful lines of inquiry. A solution to the problem of free will may still be far off, but these new developments should help free-will theorists push back against the common complaint that the free-will debate is still mired in a dialectical stalemate between “compatibilists” and “incompatibilists.”

2 Compatibility Concerns: The Arguments

From the perspective of CAP theorists, the problem of free will is just the problem of determinism, and the problem of determinism boils down to the question of whether the thesis of determinism is logically incompatible with the **classical free-will thesis**, i.e. the thesis that some ordinary human exercises free will (assuming the classical definition of ‘free will’) (Mickelson [4]). The worry, in general terms, is that a certain kind of necessity (determinism) is at odds with a kind of contingency (free will). Looked at in this way, CAP compatibility concerns are part of a family of traditional worries raised by **predeterminisms**, including not only well-known problems about determinism (Campbell and Lota [8]), God’s omniscience (foreknowledge) and the logical principle of bivalence (Finch [2]), but also eternalism (Buckareff 2019), providential determinism, and socio-economic determinism. Many of these predeterminisms involve commitments to scientific, religious, even political world views. For instance, a Catholic might be committed to providential determinism, or a Marxist to socio-economic determinism. This partly explains why some compatibility problems are worrisome to some people, but not to others. If the predeterminism is disconnected to one’s world view, it is easy to give it up. Once we consider compatibility problems broadly—as involving any number of predeterminisms in conflict with free will—it is likely each of us has a worrisome compatibility problem waiting to be revealed.

The problem of determinism remains a popular entry point to the problem of free will, but it is not the only framework which draws upon notions of luck (i.e. factors beyond our control) to raise pressing questions about the nature and existence of free will. Even if one were to show that the future is not perfectly predetermined—by God, the laws of nature, the axioms of logic, or anything else—one would not have thereby made the case for free will. Even if a world without a pre-established future must include some type of indeterminacy, it is by no means obvious which type of indeterminacy is required. This raises a new concern: perhaps the best arguments in the literature, when taken together, will support the conclusion that free will is impossible whether or not there is indeterminacy in the world and, hence, that denialism is true.

From the ancient Epicurean idea that free will might be found in the random “swerve” of Democritean atoms (Pereboom 2009, pp. 17–18) to the modern idea that free will is grounded in the (purportedly) probabilistic behavior of quantum particles (Kane 2003; Balaguer [19]), many people have argued for a tight connection between free will and causal indeterminacy. As we have seen, CAP theorists are committed to solving the problem of free will through a very particular characterization of the problem of determinism and, given their commitment to classical possibilism, classical incompatibilists are committed to a broadly libertarian interpretation of free will. However, even CAP theorists who are committed to a libertarian analysis of the ability to do otherwise respected the worry that causal indeterminacy might “hurt” one’s efforts to exercise free will. For example, van Inwagen’s “freakish demon” manipulation argument (van Inwagen 1983, pp. 130–134) was the first of the so-called *manipulation arguments* (Capes [9], Mickelson 2017) to raise serious concerns about the incompatibility of free will and *indeterminism*. The more renowned *Mind* argument raised the same concerns (Campbell and Lota [8]). (It is called the “*Mind* argument” because influential versions of it were published in the journal of that name.) The *Mind* argument “occurs in three forms” or “three closely related strands of argument that are often twisted together” (van Inwagen 1983, p. 126). All the strands begin with “a certain set of reflections on what the nature of free action must be if the incompatibilist is right,” e.g., supposing the world is causally undetermined but productive of free action.¹¹

Van Inwagen notes there are structural similarities underlying the *Mind* Argument and the Consequence Argument, the most influential argument for classical incompatibilism, suggesting that if one is sound, then so must be the other (van Inwagen 1983, pp. 147–150; Campbell and Lota [8]). In broad strokes, the Consequence Argument is a seemingly simple conditional argument: If determinism is true, then everything we will ever do is a consequence of the laws of nature and states of the world in the remote past (prior to the existence of the first human); since we have no control over the past (Wasserman [10]) or the laws (Vihvelin [14]), we have no control—in just the sense picked out by ‘free will’—over anything we do. Yet, it seems not to matter much if we replace the laws of nature with probabilistic laws. Either way, we have “precious little free will” (van Inwagen 1989, p. 405).

As a CAP theorist, van Inwagen originally presented the Consequence Argument against the backdrop of CAP’s background assumptions, which places constraints on how we interpret this argument’s premises and conclusion. For example, the original CAP-based version of the Consequence Argument (hereafter, the *Classical Consequence Argument*) was specifically an argument for classical incompatibilism. That is, the Classical Consequence Argument concludes that it is impossible to exercise the ability to do otherwise picked out by the term ‘free will’ when determinism is true, from which it follows (given the CAP assumption of classical possibilism) that a libertarian interpretation of the ability to do otherwise must be correct. According to this libertarian account, indeterministic leeway is a prerequisite for exercising the ability to do otherwise, a.k.a. free will.

Here, then, we strike a tension at the core of the CAP program. The Classical Consequence Argument concludes that classical incompatibilism is true, and above we noted that some philosophers believe that there are structural similarities between the Classical Consequence Argument and the *Mind* argument which ensure that if one of these arguments is sound, then so is the other. However, if both of these arguments are sound, it means that people cannot act freely whether determinism is true or false—in which case classical possibilism, a defining background assumption of CAP, is false.¹² The tension may indicate a problem with the assumption of classical possibilism, i.e. perhaps anthropocentric possibilism and/or the classical account of free will is incorrect (Campbell and Lota [8]). Not wanting to give up on such foundational CAP commitments, it is perhaps unsurprising that van Inwagen—an eminent CAP theorist—has responded to the apparent paradox within CAP by adopting *mysterianism*, the view that free will exists but it is a mystery (van Inwagen 1983, 1998, 2000).¹³ For those less committed to the CAP program, the best response to this tension may be less clear—though, minimally, it encourages us to explore other (neo-classical and non-classical) options.

The manipulation argument has become one of the most popular tools for exploring non-CAP approaches to the problem of free will (Capes [9]). Multiple-case manipulation arguments were already in play during the classical period, e.g. van Inwagen’s “freakish demon” argument targeted broadly libertarian accounts of free will (van Inwagen 1983, pp. 130–134) and Richard Taylor’s earlier “puppet” argument targeted compossibilist accounts of free will (Taylor 1963, p. 45). However, manipulation arguments are now used to support a wide variety of conclusions. For example, Derk Pereboom’s influential Four-Case Argument aims to establish neo-classical incompatibilism and to offer some support for the more specific source incompatibilist position that determinism-related causal factors preclude human free will by keeping people from satisfying the source condition on free will (Pereboom 2001, 2014). Other manipulation arguments are more thoroughly untethered from the CAP framework. For example, Alfred Mele’s revised Zygote Argument (Mele 2013, 2017, 2019) is distinctive insofar as it concludes to mere **impossibilism**, the relatively modest

non-explanatory claim that it is impossible for an ordinary human to act freely when determinism is true. Unlike Pereboom’s Four-Case Argument, Mele’s argument is completely silent about *why* impossibilism is true (Mickelson 2015b, 2017, 2021). Kristin Mickelson’s Master Manipulation Argument marks an even more radical departure from the classical program (Mickelson 2019a, 2019b). Like other influential non-classical arguments, such as Galen Strawson’s Basic Argument and Neil Levy’s related “Luck Pincer” (Levy 2011, Hartman [23]), the Master Manipulation Argument concludes to constitutive-luck source impossibilism—which, if true, would mean that the rival explanatory conclusion of the Four-Case Argument is *false* (Mickelson 2015b, 2017, 2019a, 2019b).¹⁴

The expansion of arguments and worries about the relationship between free will and factors beyond our control has generated new thoughts about the problem of free will, and lends force to a relatively new type of worry. With the array of views about free will now available, we can ask “Which reflects the layman’s notion of free will—and how should we respond if it turns out that there is a *conflict* between the view philosophers think is the best and the one endorsed by the folk?”. These and related worries have motivated *revisionism* about free will, the view that the correct solution to the problem of free will clashes with the folk notion of free will and/or common freedom-related practices, e.g. moral praise/blame and punishment (Vargas [13]). While revisionism raises many interesting and pressing questions about free will and metaphilosophical issues facing those who study it, the justification for revisionism will depend largely upon what our best science tells us about its empirical components (e.g. what the folk think about free will).

3 Science and Free Will

From antiquity, many philosophers have viewed the fixity of the world—whether due to gods, causal-like conditions, the principles of logic, or the like—as antithetical to the belief that humans have any control over their lives. When Newtonian physics arose, the specific challenge presented by causal determinism became especially pressing, for it quickly appeared to be foundational for a scientific view of the world. Subsequent centuries strongly reinforced the explanatory and predictive force of Newton’s mechanics with expansion of its influence into other developing sciences such as chemistry and biology, and even began to influence the development of modern psychology through Freud and later more explicitly so with Skinner and Watson’s behaviorism. On the practical side, the use of Newtonian principles became crucial to emergent technologies exhibited in the industrial revolution, providing forceful everyday evidence of their increasingly plain truth.

However, this high tide of determinism ebbed somewhat in the early 20th century with the rise of an alternative account of fundamental reality: quantum theory and the associated idea of probabilistic causation or even outright indeterminism at work in the deepest levels of reality (at least according to some interpretations). These latter interpretations brought new hope to aspiring libertarians but also raised new worries for those who believed that quantum-like indeterminacy in human nature could do nothing to aid a like account of free will, unequipped with any feature that could easily accommodate room for human control over it. In the light of these more recent scientific trends, it is not surprising then that the determinism/indeterminism debate arose and continued to strongly influence free-will theorists as informed by the constantly evolving results of scientific inquiry and emerging theories. As we related earlier, this conflict between determinism and indeterminism was philosophically sharpened and focused in the 20th century, giving rise to CAP’s emphasis on this distinction, and continues in various themes today (see Vihvelin [14] for a detailed contemporary examination of

the concept of causality at work in determinism, for example). It remains to be seen whether CAP or non-CAP perspectives along with further scientific investigation will move us closer to a satisfactory solution to the problem of free will. It is undeniable though that science has been a formative factor in the free will debate in the last century, and more recently has assumed a prominent role in the very methodology of how to conduct that debate.

In the 1980s empirical experiments such as Libet's famously began to lay ground for the still evident stand among many neuroscientists against the existence of free will (assuming that a broadly libertarian account of free will is correct) citing traceable data that unconscious predispositions for choice can be recorded even before any such choices enter the conscious realm (Libet et al. 1983; see also Waller [17]; Levy and Wright [15] examine one facet of this in terms of implicit attitudes). Since these results may be thought to favor determinism of the mind (in some sense) prior to instances of choice, then indeterminism of choice, either conscious or unconscious, would seemingly be ruled out (but see Woodward [16]). Libertarianism thus appears to be completely knocked out of the realm of plausibility (though many following Libet simply then do not make argumentative room for the possibility of a compatibilist/compossibilist view of free will, thus revealing their bias that some broadly libertarian account of free will must be correct; see Cova [18] for discussion). However, many have pointed out that this is a rush to judgment given the uncertainty of what the data really reveals as against several plausible alternatives of how metaphysically choice may work moving from unconscious sources into the arena of how conscious choices are made (Robichaud [21] examines one important aspect of this).

The rise of neuroscience in the latter part of the 20th century also now plays an important—and some say indispensable—role in understanding how free will and action theory issues sort out against the background of studies about the brain and mental behavior ((Waller [17]); on one extreme end Penrose 1989 argues that a quantum theory of brain activity may solve the free will problem and in a way favorable to libertarianism, but see Boolos et al. 1990 for criticism). Indeed, some advocates of libertarianism have insisted that a careful examination of the science of the brain supports that view and speaks against a Libet-style conclusion of his own studies (Balaguer [19]; see also Kane 2003). As we better understand the details of our mental lives in scientific terms, we may discover at least important clues about how to better interpret any role that freedom and responsibility might play out with respect to those findings.

Aside from the determinism/indeterminism debate, the most direct empirical trend of the 21st century involving the free will problem has been in the rise of experimental philosophy—usually termed “X-Phi” (see Nahmias et al. 2005 and Nichols and Knobe 2007 for example; for metacriticism see Cova et al. 2018). The motivation for X-Phi is rooted in familiar claims in previous free will literature (especially the CAP-based distinction between compatibilism and incompatibilism) about the beliefs and attitudes of “the folk.”¹⁵ X-Phi developed in part to inform such claims with real data—statistical surveys that posed specific sets of questions to groups of individuals in order to ascertain real-world beliefs and attitudes about matters of freedom and free will. The idea was that if one could obtain real world data about the intuitions of large groups of people about specific free will-related scenarios then one could tabulate in a comprehensive way overall views that then could factor into free will debates, thus eliminating pure speculation about how “the folk” were disposed to talk in favor of a scientific basis for such claims. In addition, these empirical methods have been extended to include methods testing for psychological factors such as implicit bias, which subconsciously could influence conscious decisions of a free will nature (Levy and Wright [15]). While the significance of this overall avenue of inquiry is still controversial, there is little doubt that these empirical methods will have continuing influence in the ways that forthcoming debates on free will are framed.

4 Moral Responsibility

While the relationship between free will and moral responsibility has always been part of the free-will debate, the latter has become even more prominent as CAP's influence has waned. When the classical (ability to do otherwise) characterization of free will was challenged, philosophers generally agreed that a more inclusive definition of 'free will' was needed for purposes of framing the problem of free will and the problem of determinism. The new definition needed to avoid any details that might be considered question-begging (e.g. by presuming that a classical rather than source account of free will is correct, or vice versa), yet it also needed to be adequately precise to pick out a distinct phenomenon as the topic of debate (lest the free-will discourse devolve into an empty verbal dispute). In response, many philosophers have adopted the practice of using 'free will' to refer to the necessary *control condition*—as opposed to the necessary *epistemic condition* (Robichaud [21])—on moral responsibility, where the latter is understood in the backward-looking, non-consequentialist, type of responsibility associated with *basic desert* (McCormick [24]).

Whether or not one approves of the move towards identifying free will with the type of control required for basic-desert moral responsibility, the moral-responsibility turn in the free-will literature has had its benefits. While moral responsibility is interesting in its own right, the neo-classical practice of fixing the referent of 'free will' in terms of moral responsibility has helped us to approach the problem of free will in new ways and encouraged us to reconsider what the free-will debate is and/or should be about (White [31]). For example, it is commonly agreed that free will is a type of *control* that one exercises in the performance of an action, and that anyone with free will would have, at minimum, the type of control required to make a person praiseworthy and/or blameworthy for their morally-valenced actions. As such, free will seems required to make a person an apt target of the *moral emotions* (Ekstrom [22]) and familiar practices of praise and blame. As such, settling what type of control (if any) is really needed for these things may help us to get a better grasp on what a viable solution to the problem of free will must look like. For example, many free-will theorists have been skeptical of the proposal that 'free will' picks out (or should pick out) the type of control required for ultimate "heaven-and-hell" moral responsibility (Strawson 1994, Adams [7]), a type of responsibility implicit in the belief that God will ensure that humans receive their just deserts, e.g. being tormented in hell in the afterlife. Not only does such *ultimate control* seem to be metaphysically impossible or even incoherent (e.g. Strawson 1994, p. 8; van Inwagen 1998, Mickelson 2019b), but some hold that a comparatively modest type of control—perhaps even more modest than basic-desert responsibility (if there is a difference between the two)—would be enough to support our current moral practices of praise/blame (McCormick [24]), forgiveness (Ekstrom [22]), and reward/punishment. If this is right, then perhaps the idea that free will is intimately connected with some type of ultimate or basic-desert control is mistaken. However, if such practices and policies are justified only if we are at least basic-desert responsible for our actions, then free-will denialism would seem to imply that the time has come to revise these and closely related practices and policies (or at least the justification for them), such as legal policies which recommend harsh punishments for criminal behavior (Ekstrom [22], K. Levy [25], Pereboom [29]).

Among the more recently developed moral-responsibility approaches to the problem of free will are those which focus on the moral agency of the mentally disabled and young children (Griffith [20]) and those which draw upon the well-established literature on the *paradox of moral luck* (Hartman [23]). Since the paradox of moral luck emerged during the classical period of the free-will debate (Williams and Nagel 1976; see also Nagel 1986), it is to be expected that

moral-luck theorists have typically assumed the CAP view of the problem of determinism, i.e. that it is a narrow debate about “antecedent causal luck”. However, the recent convergence of the basic vocabulary between the two literatures—especially the language of *control* and *luck*—has highlighted hitherto overlooked similarities between the two problems. As the cross-pollination of these established literatures increases, we can expect more critical pressure to fall upon the CAP assumption that the problem of determinism is fundamentally a problem of causal luck—as opposed to, say, a problem of constitutive luck (Mickelson 2019b). Future work which explores these non-classical avenues of thought may prove equally useful to philosophers working on the problem of free will and to those philosophers who are interested in free will only insofar as it is related to moral and legal responsibility.

5 The Future

The wisdom of speculation about the future of anything has considerable history against it. From the supreme confidence in Newtonian physics prior to Einstein and Planck to the declaration that World War I was so horrific as assuredly to constitute “the war to end all wars”, there are countless examples of the retrospective frivolity of predicting the future that seem to undercut the wisdom of even the attempt to do so. However, just as the role of hypothesis is central to the work of much science, and has proved its merits time and again even though failures vastly outnumber successes, we believe that some prognostication about the future of free will and action theory might yield some parallel advantage. This is how we approach such an effort in this volume—not only directly trying to predict how things might go in these and related areas (Mele [30], Tierney [27]), but also emphasizing present areas of investigation that might prove much more fruitful in the future.

The major future trend we see in several contributions in this volume is an emphasis on the role of empirical methods in contributing to or even guiding the dialogues on free will and action theory, as we noted above concerning the rise of X-Phi and the increasing influence of neuroscience. Another facet of this empirical dimension to the debates is that there appears to be an increased emphasis on the phenomenology of choice (Deery and Nahmias [26], Robichaud [21], Woodward [16]; also see Mele [30]). How such an introspective factor argumentatively plays off against more traditional empirical treatments of psychological data requires much more investigation.

A separate trend rooted in empiricism is that of a pragmatic approach to the free will problem, which although implicitly present in the literature since at least P.F. Strawson’s influential “Freedom and Resentment” (Strawson 1962), has not been overtly promoted as a dominant theme. Revisionism (Vargas [13]) incorporates some trace of this in its relativizing the adequacy of a concept of free will to its overall workability at any given time (see Ekstrom [22] on this as well). Illusionism (Smilansky [12]) is partly pragmatic by conceding the falsity of libertarianism yet arguing that we need such a concept in moral and legal terms in order to best work as societies (see Levy [25] on how free will is incorporated in matters of legality). In this volume it is argued as well that pragmatism yields the best approach to defining key concepts such as determinism (White [3]), and perhaps is the best overall approach to the entire free will problem (White [31]).

Beyond considerations of the empirical, it appears clear that forms of free will denialism, skepticism, and even to an extent illusionism will expand in influence (Hartman [23], Pereboom [29], Smilansky [12]; also see Vilhauer 2012 and Mele [30] who offer varieties of epistemic skepticism about free will). Though some routes to skepticism, denialism (especially

Pereboom's), and illusionism are extensions of the CAP program, others are firmly outside the CAP tradition (e.g. Galen Stawson 1986, Levy 2011, Mickelson 2019b; see also Mickelson [4]). These non-classical approaches have put considerable pressure on the CAP assumption of classical possibilism, and we believe especially that these forms of skepticism and denialism will increase in influence. Non-classical explorations may well have an influence on future developments in X-Phi as well, leading to better inquiries informed by considerations of the roles of luck in our choices and moral lives and perhaps even leading us to see that intuitions favoring impossibilism are more widespread than currently assumed. Such inquiries could lead to an increased pragmatic concern with reforming our more formal and legal blaming practices in society.

Of course, familiar philosophical approaches in the tradition of CAP or in direct/indirect criticism of it will certainly also maintain significant influence in much or most of the literature (in this volume: Campbell and Lota [8], Mickelson [4], Balaguer [19], Adams [7], Nichols [28], Finch [2], Leftow [11], Mele [30], Vicens [5], Pereboom [29], White [31] for example) and it is clear that this is an important part of moving the field forward by the continued reliance on the time-honored methods of logically-constrained speculation. After all, even in science the source of hypothesis is always the rigorous application of the inventive prowess of the human mind to intriguing problems.

Glossary

Anthropocentric Impossibilism: The view that it is metaphysically impossible for an ordinary human to exercise free will. Anthropocentric impossibilism entails **free-will denialism**, but it does not entail **impossibilism**. **Hard incompatibilism** is a species of anthropocentric impossibilism, though it is *not* a species of impossibilism.

Anthropocentric Possibilism: The view that it is metaphysically possible for an ordinary human to exercise free will. Anthropocentric possibilism entails that **anthropocentric impossibilism** is false and that **possibilism** is true; it is silent vis-a-vis the truth-value of the **free-will thesis** and **free-willism**.

Classical (a.k.a Leeway) Account of Free Will (or the classical definition of 'free will'): Free will is an ability to do otherwise; typically contrasted with *source* accounts of free will. Within CAP, the classical account of free will was assumed to be true, leaving open the debate between **classical compatibilists** and **classical incompatibilists** over which interpretation of the ability to do otherwise is correct.

Classical Analytic Paradigm (CAP): The dominant research paradigm during the classical period (c. 1965–1985) of the free-will debate. The terms 'compatibilism' and 'incompatibilism' were coined for use within CAP, and the background assumptions of CAP play an essential role in justifying the familiar CAP narrative that these two terms named the only two viable candidate solutions to the **classical compatibility problem**.

Classical Compatibilism: The CAP-based view that the **classical account of free will** is correct and that it is metaphysically possible for an ordinary human to exercise free will (where 'free will' refers to an ability to do otherwise) when **traditional determinism** is true, i.e. necessarily, determinism is logically compatible with the **classical free-will thesis** (Mickelson [4]). The "classical" qualifier signals that the classical definition of 'free will' is used in stating the view. The term 'compatibilism' was coined (in the 1960s) as a name for this view.

Classical Compatibility Problem: According to CAP theorists (i.e. philosophers who endorse and work within CAP), the **problem of determinism** boils down to the challenge of settling whether **classical compatibilism** or **classical incompatibilism** is true. Given that **classical possibilism** is a background assumption of CAP, all classical compatibilists and classical incompatibilists were **anthropocentric possibilists**.

Classical Free-Will Thesis: The thesis that free will is (or requires) an ability to do otherwise (i.e. the **classical account of free will** is correct) and some ordinary human exercises free will; put another way, the thesis that an ordinary human exercises free will, where 'free will' refers to an ability to do otherwise.

Classical Incompatibilism: The CAP-based view that **the classical account of free will** is correct and it is metaphysically impossible for an ordinary human to exercise free will when **traditional determinism** is true because determinism-related factors preclude the type of indeterministic leeway that an exercise of free will requires, i.e. necessarily, traditional determinism is logically incompatible with the **classical free-willism** (Mickelson [4]). The “classical” qualifier in the name signals that the classical definition of ‘free will’ is used in stating the view. The term ‘incompatibilism’ was coined (in the 1960s) as a name for this view. (If classical incompatibilism is true, then so is **neo-classical incompatibilism**, but not vice versa.) The Classical Consequence Argument (i.e. the Consequence Argument, as originally presented against the background of CAP) concludes to classical incompatibilism.

Classical Possibilism: the conjunction of two views: (1) free will is (some kind of) an ability to do otherwise, i.e. the so-called **classical account of free will** is correct, and (2) **anthropocentric possibilism**.

Compossibilism: The view that it is metaphysically possible for an ordinary human to exercise free will in a world at which traditional determinism is true; the conjunction of determinism and the **free-will thesis** is metaphysically possibly true.

Constitutive-luck Source Impossibilism: The view that it is metaphysically impossible for anyone (i.e. any metaphysically possible being) to exercise free will because constitutive luck—as opposed to, say, causal luck—prevents people from satisfying the necessary source condition on free will. Galen Strawson’s Basic Argument (Strawson 1994, 2011) and Kristin Mickelson’s Master Manipulation Argument (Mickelson 2021) each conclude to this view.

Free-will Denialism: One of two basic solutions to the problem of free will (the other is **free-willism**). Denialism is the view that no (ordinary human) has free will, i.e. the view that **the free-will thesis** is false. Hard determinism is a common route to denialism; all arguments for **anthropocentric impossibilism** and **impossibilism** are (*a fortiori*) arguments for denialism.

Free-will Thesis: The thesis that an ordinary human exercises free will (where ‘free will’ is neutral between classical and non-classical accounts of free will, e.g. by fixing the referent of ‘free will’ to the control condition on basic-desert moral responsibility). Compare to the **classical free-will thesis**.

Free-willism: One of two basic solutions to the problem of free will (the other is **free-will denialism**). The view that some ordinary human has free will, i.e. the view that the **free-will thesis** is true. Libertarianism and soft determinism are common species of free-willism.

Hard incompatibilism: A species of **anthropocentric impossibilism** which claims that it is impossible for an ordinary human to exercise free will on the grounds that, whether determinism is true or false, some kind of causal luck (i.e. causal factors beyond one’s control) would keep a normal human from satisfying some necessary condition on free will. The “hard” in the name signals that the view entails **free-will denialism**. Notably, hard incompatibilism is a species of **anthropocentric impossibilism** but not (unqualified) **impossibilism**.

Hard Source Incompatibilism: The view that **hard incompatibilism** is true, and the necessary condition which an ordinary human cannot satisfy when determinism is true is a *source* condition and not a classical ability-to-do-otherwise (a.k.a. leeway) condition.

Impossibilism: The unqualified view that it is metaphysically impossible for anyone (i.e. any metaphysically possible being, even God) to exercise free will. Impossibilism entails **denialism**. Galen Strawson’s “Basic Argument” (Strawson 1994, 2011) and Kristin Mickelson’s “Master Manipulation Argument” (Mickelson 2021) conclude to impossibilism. (Notably, **hard incompatibilism** is not an impossibilist view.)

Incompatibilism: The term ‘incompatibilism’ has become an umbrella term and currently has no standard meaning; the same is true of many phrases commonly associated with this term, e.g. “free will is incompatible with determinism”. The term is currently used to refer to **impossibilism**, **classical incompatibilism**, **neo-classical incompatibilism**, **anthropocentric impossibilism**, **impossibilism**, and many other views (e.g. see endnote 10). (The same is true, *mutatis mutandis*, of the term ‘compatibilism’ and the ambiguous phrases commonly used to define it).

Impossibilism: The view that it is metaphysically impossible for an ordinary human to exercise free will in a world (or universe) at which **traditional determinism** is true; alternatively, the view

that the material conditional “If traditional determinism is true, then the **free-will thesis** is false” is necessarily true (i.e. true in all possible worlds) (for more detail, see Mickelson [4]). Some philosophers now use ‘incompatibilism’ to refer narrowly to impossibilism (e.g. Mele [30], Capes [9]; see endnote 10 for discussion). Within CAP, any argument for impossibilism was also (given CAP background assumptions) an argument for classical incompatibilism; outside of CAP, the inference from impossibilism to classical incompatibilism or neo-classical incompatibilism is a fallacious *cum hoc, ergo propter hoc* (“with this, therefore on account of/because of this”) inference (see Mickelson [4] and 2021). Alfred Mele’s revised Zygote Argument (Mele 2013, 2017, 2019; Mickelson 2015b) is an example of an argument for mere impossibilism.

Neo-Classical Compatibilism: The view that it is metaphysically possible for an ordinary human to exercise free will when **traditional determinism** is true, and traditional determinism does not stand in any antagonistic relevance relation to free will (where ‘free will’ is neo-classically characterized in a way that is neutral between classical and non-classical accounts of free will). The difference between **neo-classical compatibilism** and **classical compatibilism** has to do with how the free-will relatum of each view is fleshed out: the latter assumes that the classical account of free will is correct but the former does not.

Neo-Classical Incompatibilism: The view that it is metaphysically impossible for an ordinary human to exercise free will (where ‘free will’ is neutral between classical and non-classical accounts of free will) when **traditional determinism** is true because there is a type of antagonistic relevance relation between free will and determinism-related factors; alternatively: necessarily, determinism is *logically incompatible* with the **free-will thesis** (Mickelson [4]). The difference between neo-classical incompatibilism and **classical incompatibilism** is in the free-will relatum, namely that the latter presumes that the **classical account of free will** is correct and the former does not. Derk Pereboom’s Four-Case Argument (Pereboom 2001, 2014) is a famous argument for neo-classical incompatibilism. See endnote 10 for further discussion.

Predeterminism: Predeterminism is a trans-temporal (past-to-future) form of determining (fixing, setting, etc.) of events and/or the truth-values of propositions, and its forms of determination include principles like bivalence, divine foreknowledge and providence, **traditional determinism**, eternalism, and socio-economic determinism.

Problem of Determinism: A loose collection of disagreements about how to spell out the relation and relata of the template-question “Does free will stand in relation R to determinism: *yes* or *no*?”, and to explain how (if at all) asking and answering one or more instances of this template-question would help us to solve the **problem of free will**. While CAP theorists treated the **problem of determinism** narrowly as a problem of causal luck (i.e. the challenge of settling whether determinism-related causal and/or nomological factors preclude human free will), non-CAP theorists have suggested alternative characterizations (e.g. that determinism scenarios, like manipulation cases, sensitize us to threats posed by constitutive luck).

Problem of Free Will: The debate over the nature and existence of free will. In generic terms, the two basic solutions to the problem of free will are **free-willism** and **free-will denialism**. A complete solution to the problem of free will—and hence a complete statement of free-willism or free-will denialism—must spell out what free will is, e.g. by proposing a set of individually necessary and jointly sufficient conditions for acting freely.

Traditional Determinism: The doctrine that one unique future (relative to any arbitrary time *t*) is the inevitable result of the naturalistic factors which account for the evolution of the physical world over time (e.g. certain future-fixing causal and/or nomological relations between events in the past and events in the future). Notably, traditional determinism is not open to a so-called “broadly Humean” interpretation (Beebe and Mele 2002), for it affirms the presence of just the sort of necessity-in-nature that broadly Humean accounts of causation/laws of nature (by definition) reject; the doctrine known as “Humean determinism” is a species of **traditional indeterminism**. In this chapter, ‘determinism’ refers to traditional determinism and ‘indeterminism’ refers to traditional indeterminism unless otherwise stated. This is just one of many doctrines which goes by the name ‘determinism.’ (See also endnote 6 and White [3] for further discussion.)

Traditional indeterminism: The thesis that **traditional determinism** is false.

Notes

- 1 The term 'free-willist' has been used as an alternative name for the free-will libertarian (e.g. William James 1921, A.J. Ayer 1968, and Robert Kane 1996) and the term 'free-willism' is commonly associated with the theological position of Armenianism. We do not follow such usage here. The term 'free-will skepticism' is sometimes used to refer to denialism and/or an epistemic position about what we are justified in believing vis-a-vis the truth of denialism (e.g. McKenna and Pereboom 2016, p. 32); we editors prefer to restrict 'denialism' to a claim about the existence of free will and to restrict the term 'skepticism' to epistemic positions (e.g. the view that *we are justified in believing* that denialism is true and/or the more modest view that *we are not justified in believing* that free-willism is true).
- 2 No name was assigned to someone who embraced the conjunction of denialism and indeterminism, i.e. someone who answered the two questions above "no" and "no" (though 'hard indeterminist' would be apt.)
- 3 The terms 'compatibilism' and 'incompatibilism' were, by all appearances, coined by Keith Lehrer in the 1960s and were greatly popularized by Peter van Inwagen, especially van Inwagen 1983.
- 4 A *broadly libertarian account* of free will is one which proposes that it is at least metaphysically possible for someone to act freely, but includes at least one necessary condition which is metaphysically impossible to satisfy when determinism is true (e.g. Vicens [5], Adams [7], Smilansky [12], Balaguer [19]). Notably, one may adopt a broadly libertarian account of free will without accepting free-will libertarianism or anthropocentric possibilism (e.g. Pereboom 2014; Mickelson "Hard Times for Hard Incompatibilism", ms.).
- 5 Another notable feature of CAP is that CAP theorists typically focused on logical relationships between propositions rather than metaphysical relationships between non-propositional phenomena. For example, van Inwagen introduced a logical entailment thesis to capture the traditional metaphysical doctrine of determinism (see van Inwagen 1990, p. 277 for helpful diagrams), and used this entailment thesis as a proxy for traditional determinism in logic-text proofs which aimed to demonstrate that a *strict logical inconsistency* relation holds between between determinism and the **classical free-will thesis** (e.g. van Inwagen 1983; see also Mickelson [4]).
- 6 The traditional doctrine of determinism is interesting, in part, because it provides the limiting-case doctrine for minimal actual-sequence leeway, i.e. it states that there is literally zero indeterministic leeway (of any kind) in the world. Assuming determinism, not even God could intervene to prevent the "determined" future from coming to pass (e.g. van Inwagen 1990, p. 277, Sehon 2011, Mickelson 2012). As such, determinism was a useful tool for exploring free will as it was classically characterized. However, philosophers have provided interesting reasons for thinking that there are other—at least equally good or better—ways of defining 'determinism' vis-à-vis the problem of free will (e.g. Dennett 2003, White [3]).
- 7 The neo-classical compatibility problem is evident when philosophers frame the problem of determinism as a debate about whether determinism (determinism-related factors, deterministic causal luck, or the like) is a threat to free will because it keeps people from being able to act otherwise and/or because it keeps people from being an adequate source of their own actions (Mickelson [4]). When philosophers adopt this neo-classical framework, they struggle to classify views—such as **constitutive-luck source impossibilism** (discussed below)—which insist that it is impossible to act freely when determinism is true even though determinism itself is entirely irrelevant to free will. This is notable, given that worries about constitutive luck have been present in discussions of the problem of determinism since its inception, as can be seen in the surviving records of the debates between the Stoics and their critics (see Pereboom 2009, Ch. 2).
- 8 Semi-compatibilism is the result of another notable attempt to re-orient the problem of determinism around the specific type of control required for moral responsibility in order to evade the narrow use of 'free will' established by CAP theorists (e.g. Fischer 1994; Fischer and Ravizza 2006; Fischer and Ravizza 1998). While semi-compatibilists agree that determinism-related factors may undermine some types of control (e.g. "regulative control"), they insist that such factors do not undermine the control required for moral responsibility. The semi-compatibilist sidesteps a direct challenge to

CAP theorists about the meaning of ‘free will’ by taking no stand on what this term does (or should) mean—except to say that *if* a person insists on using the classical definition of ‘free will’ and it turns out that determinism-related factors preclude one’s *ability to do otherwise* (as classical incompatibilists claim), this result would not establish that determinism-related factors preclude moral responsibility; it would, rather, show that free will (i.e. the ability to do otherwise) is not required for moral responsibility. As such, semi-compatibilism is distinct from neo-classical forms of compatibilism which fix the referent of ‘free will’ to the control condition on moral responsibility only insofar as the latter take a stand on what free will is (and what ‘free will’ means) while the semi-compatibilists do not.

- 9 Although Pereboom is not an impossibilist, he is an anthropocentric impossibilist because he denies that a being who has the properties of an ordinary human (i.e. someone like us, as we are in the actual world) can satisfy the “law-overriding” source condition he forwards as part of his broadly libertarian account of free will. That is, hard incompatibilism is a type of anthropocentric impossibilism, but is not a type of impossibilism (for discussion, see Mickelson “Hard Times for Hard Incompatibilism”, ms.).
- 10 As philosophers moved away from the original CAP-based definitions of ‘compatibilism’ and ‘incompatibilism’, they retrofitted the qualifier “classical” to these terms as a way of marking that departure. (The terms ‘classical compatibilism’ and ‘classical incompatibilism’ are also applied to views held by pre-CAP philosophers, which leads to complications we cannot address here.) Adding such qualifiers is one way to show due respect for the methodological principle that philosophers may define their jargon however they like while keeping tabs on the dialectically significant variations currently in use. Following suit, we have applied the qualifier “neo-classical” to single out the initial successors to the classical characterizations. Expanding this tracking device, we wish to identify a few additional recharacterizations which may be of interest to the reader.

As already noted, *neo-classical incompatibilism* has two defining tenets, namely impossibilism and a positive explanatory thesis which states roughly that impossibilism is true because determinism (determinism-related causal/nomological factors) deprive ordinary humans of free will; *neo-classical compatibilism* is also a two-tenet view, one tenet negates the negative thesis of neo-classical incompatibilism and the other negates its positive thesis (which means that neo-classical compatibilism is not equivalent to mere compossibilism (Mickelson 2012, 2015a)). Some philosophers have proposed that we use ‘incompatibilism’ to denote only the positive thesis of neo-classical incompatibilism and ‘compatibilism’ to name its negation (Levy 2011: p. 1, n. 1; Mickelson 2015b); let these be *anti-classical incompatibilism* and *anti-classical compatibilism*, respectively. Assuming this anti-classical revision, impossibilism is not a defining tenet of incompatibilism but remains a corollary, so the anti-classical redefinition of ‘incompatibilism’ leaves the term’s earlier neo-classical meaning largely intact. However, anyone who rejects anti-classical incompatibilism qualifies as an anti-classical compatibilist, which means that some *impossibilists* qualify as compatibilists on this anti-classical taxonomy. Anti-classical theorists consider this a feature rather than a bug, for it highlights that some philosophers argue for the negative thesis of neo-classical incompatibilism but against its positive thesis—a position that is not supposed to be available according to popular CAP-based narratives. (Kadri Vihvelin aims to achieve a similar goal via alternative terminological revisions, roughly: keep compossibilism as a defining tenet of ‘compatibilism’, redefine ‘incompatibilism’ to pick out the conjunction of impossibilism and anthropocentric possibilism, and add ‘impossibilism’ to refer to anthropocentric impossibilism (e.g. Vihvelin 2008, 2013). A downside of this “tripartite taxonomy” is that anthropocentric impossibilists (e.g. hard incompatibilists) cannot be classified as incompatibilists even when they embrace both tenets of neo-classical incompatibilism (e.g. Vihvelin 2013: p.242, n. 5; Mickelson 2015a)). Other philosophers now use ‘compatibilism’ and ‘incompatibilism’ as their preferred labels for compossibilism and impossibilism (e.g. Mele [30] and 2017: p. 6, n. 4; Capes [20]); to track this usage, let these be *post-classical compatibilism* and *post-classical incompatibilism*, respectively. These post-classical revisions bring back a bipartite taxonomy of (in)compatibilism by rejecting—fruitfully, according to post-classical theorists—more complicated taxonomies which treat the traditional dispute between anti-classical compatibilists and anti-classical incompatibilists as a fundamental point of divide in the contemporary free-will debate. A purported upside of the post-classical tax-

onomy is that impossibilists cannot be compatibilists (since post-classical compatibilists are compossibilists); a downside is that anti-classical compatibilists and anti-classical incompatibilists are lumped into one motley “anti-compatibilist” category. Hybrid recharacterizations are also found in the literature, e.g. using ‘incompatibilism’ to denote neo-classical incompatibilism but ‘compatibilism’ to denote mere compossibilism (see Mickelson 2021 for discussion). Despite appearances, the latter hybrid does not yield a genuine bipartite taxonomy, for (assuming these hybrid definitions) it may be that compatibilism and incompatibilism are both false and some third view—unnamed by the hybrid theorist—is true (e.g. constitutive-luck impossibilism).

With the above distinctions in hand, readers are better prepared to spot the common practice of technically defining ‘incompatibilism’ in one way while using it in another (e.g. McKenna and Pereboom 2016: pp. 30 and 151; Sartorio 2016: pp. 147 and 157) and to track fundamental differences between famous “arguments for incompatibilism”. For example, Pereboom’s Four-Case Argument (Pereboom 2001, 2014) concludes to neo-classical incompatibilism (Pereboom [29, n.5]), relying upon a slippery-slope argument to support post-classical incompatibilism and a modest best-explanation argument to support anti-classical incompatibilism; Alfred Mele’s original Zygote Argument (Mele 2006) is invalid because its premises support mere post-classical incompatibilism but its conclusion is a statement of anti-classical incompatibilism (and/or neo-classical incompatibilism) (Mickelson 2015b); Mele’s revised Zygote Argument (e.g. Mele 2013, 2017, 2019) is an argument for post-classical incompatibilism (a.k.a. impossibilism) but it is not an argument for anti-classical incompatibilism (and hence is not an argument for neo-classical incompatibilism); Kristin Mickelson’s Master Manipulation Argument—like Galen Strawson’s Basic Argument (Strawson 1994)—concludes to impossibilism via reasoning which implies that post-classical incompatibilism (a.k.a. impossibilism) is a true but metaphysically trivial position and that anti-classical incompatibilism (hence neo-classical incompatibilism) is false (e.g. Mickelson 2015b, 2019a, 2019b, “Hard Times for Hard Incompatibilism,” ms.). Again, the novel qualifiers we have applied to the term ‘incompatibilism’ here are merely a rhetorical device for tracking the different views currently called by name ‘incompatibilism’; whether philosophers should continue to use a single term in such disparate ways is another matter. Indeed, Mickelson argues that non-CAP theorists should sidestep the project of rehabilitating jargon from a research paradigm they reject and instead embrace new ways of talking about the fundamental divides in the contemporary debate (e.g. Mickelson [4]). Readers are advised to keep such differences and debates in mind as they read the chapters in this volume, and are invited to consider their own preferred solution to these jargon/taxonomy problems.

- 11 Since all the strands are critical of libertarianism, a CAP theorist may interpret these arguments as lending support to classical compatibilism.
- 12 Notably, other interesting problems arise when we untether the Consequence Argument from CAP, e.g. it is unclear that the argument still pinpoints determinism as a threat to free will (e.g. Campbell 2007).
- 13 Van Inwagen finds some logical space in the possibility of imminent or agent causation but this just raises further puzzles (van Inwagen 1983: 151–52).
- 14 Notably, this means that if the conclusion of Mickelson’s Master Manipulation Argument is true, Pereboom’s hard incompatibilism is also false. Readers should note that interesting dialectical points like this one are often obscured by the common classically-driven practice of lumping all manipulation arguments together under the label “arguments for incompatibilism.” We advise the reader to look carefully at the stated conclusion of any given manipulation argument in order to decide whether the argument aims to support mere impossibilism, a type of classical or neo-classical incompatibilism, or a type of impossibilism which entails that impossibilism is true but classical and neo-classical forms of incompatibilism are false.
- 15 If, as we claim, the background assumptions of much X-Phi inquiry is within the tradition of CAP, then many issues, such as the role that non-causal types of luck (e.g. constitutive luck) may play in deterministic or indeterministic scenarios, are completely left out of the picture, and this may well skew the subjects’ responses in errant ways. Perhaps X-Phi studies might better reflect non-CAP concerns in the future? (For positive signs of movement in that direction, see Cova 2022.)

Bibliography

- Ayer, A.J. (1968). *The Origins of Pragmatism*. London: Macmillan.
- Beebe, H. and Mele, A. (2002). Humean compatibilism. *Mind* 111: 201–223.
- Bobzien, S. (1998). *Determinism and Freedom in Stoic Philosophy*. Oxford: Oxford University Press.
- Bobzien, S. (2021). *Determinism, Freedom, & Moral Responsibility: Essays in Ancient Philosophy*. Oxford: Oxford University Press.
- Boolos, G. et al. (1990). An open peer commentary on the emperor's new mind. *Behavioral and Brain Sciences* 13 (4): 655.
- Buckareff, A.A. (2019). Time, leeway, and the laws of nature: why humean compatibilists cannot be eternalists. *Metaphysica* 20 (1): 51–71.
- Campbell, J.K. (2007). Free will and the necessity of the past. *Analysis* 67 (2): 105–111.
- Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.
- Cova, F. (2022). It was all a cruel angel's thesis from the start: folk intuitions about zygote cases do not support the zygote argument. In: *Advances in Experimental Philosophy of Free Will and Responsibility* (ed. T. Nadelhoffer and A. Monroe). London: Bloomsbury Academic.
- Cova, F. et al. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology* 12: 9–44.
- Dennett, D. (2003). *Freedom Evolves*. London: Penguin Books.
- Fischer, J.M. (1994). *The Metaphysics of Free Will: An Essay on Control*. Oxford: Blackwell.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
- Fischer, J.M. and Ravizza, M. (2006). *My Way: Essays on Moral Responsibility*. New York: Oxford University Press.
- Frankfurt, H.G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66 (23): 829–839.
- Kane, R. (2003). Free will: new directions for an ancient problem. In: *Blackwell Readings In Philosophy: Free Will*, 3e. Oxford: Wiley-Blackwell. 222–248.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.
- Lehrer, K. (1990). *Metamind*. Oxford: Clarendon Press.
- Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. New York: Oxford University Press.
- Libet, B. et al. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): the unconscious initiation of a freely voluntary act. *Brain* 106 (3): 623–642.
- McKenna, M. (2010). Whose argumentative burden, which incompatibilist arguments?—Getting the dialectic right. *Australasian Journal of Philosophy* 88 (3): 429–443.
- McKenna, M. and Pereboom, D. (2016). *Free Will: A Contemporary Introduction*. New York: Routledge.
- Mele, A. (2005). A critique of pereboom's 'four-case argument' for incompatibilism. *Analysis* 65 (1): 75–80.
- Mele, A. (2006). *Free Will and Luck*. New York: Oxford University Press.
- Mele, A. (2013). Manipulation, moral responsibility, and bullet biting. *Journal of Ethics* 17 (3): 167–184.
- Mele, A. (2017). *Aspects of Agency: Decisions, Abilities, Explanations, and Free Will*. Oxford: Oxford University Press.
- Mele, A. (2019). *Manipulated Agents: A Window to Moral Responsibility*. Oxford: Oxford University Press.
- Mickelson, K.M. (a.k.a. K. Demetriou). (2012). *Free Will Fundamentals: Agency, Determinism, and (In)compatibility*. [Dissertation, University of Colorado, Boulder]. https://scholar.colorado.edu/concern/graduate_thesis_or_dissertations/g732d9110.
- Mickelson, K.M. (a.k.a. K. Demetriou). (2015a). A critique of Vihvelin's three-fold classification. *Canadian Journal of Philosophy* 45 (1): 85–99.
- Mickelson, K.M. (a.k.a. K. Demetriou). (2015b). The zygote argument is invalid—now what? *Philosophical Studies* 172 (11): 2911–2929.
- Mickelson, K.M. (a.k.a. K. Demetriou). (2017). The manipulation argument. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy). New York: Routledge. 166–178

- Mickelson, K.M. (a.k.a. K. Demetriou). (2019a). Free will, self-creation, and the paradox of moral luck. *Midwest Studies in Philosophy* 43 (1): 224–256.
- Mickelson, K.M. (a.k.a. K. Demetriou) (2019b). The problem of free will and determinism: an abductive approach. *Social Philosophy & Policy* 36 (1): 154–172.
- Mickelson, K.M. (a.k.a. K. Demetriou). (2021). The zygote argument is still invalid: so what? *Philosophia* 49 (2): 705–722.
- Nagel, T. (1976). Moral luck II. *Proceedings of the Aristotelian Society, Supplementary Volumes* 50: 137–151.
- Nagel, T. (1986). *The View from Nowhere*. Oxford: Oxford University Press.
- Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2005). Surveying freedom: folk intuitions about free will and moral responsibility. *Philosophical Psychology* 18 (5): 561–584.
- Nichols, S. and Knobe, J. (2007). Moral responsibility and determinism: the cognitive science of folk intuitions. *Noûs* 41 (4): 663–685.
- Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.
- Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2009). *Free Will*, 2e (ed. D. Pereboom). Indianapolis: Hackett Publishing Company.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.
- Sartorio, C. (2016). *Causation and Free Will*. Oxford: Oxford University Press.
- Sehon, S. (2010). A flawed conception of determinism in the consequence argument. *Analysis* 71: 30–38.
- Shabo, S. (2011). “What must a proof of incompatibilism prove?” *Philosophical Studies* 154: 361–371.
- Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies* 75 (1/2): 5–24.
- Strawson, G. (2011). Free will. In: *Routledge Encyclopedia of Philosophy* (ed. E. Craig). London: Routledge. <https://www.rep.routledge.com/articles/thematic/free-will/v-1/sections/pessimism>.
- Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy* 48: 1–25.
- Swenson, P. (2016). Ability, foreknowledge, and explanatory dependence. *Australasian Journal of Philosophy* 94: 658–71.
- Taylor, R. (1963). *Metaphysics*, 1e. Englewood Cliffs: Prentice-Hall.
- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Oxford University Press.
- van Inwagen, P. (1989). When is the will free? *Philosophical Perspectives* 3: 399–422.
- van Inwagen, P. (1998). The mystery of metaphysical freedom. In: *Metaphysics: The Big Questions* (ed. P. van Inwagen and D.W. Zimmerman). Oxford: Blackwell Readings in Philosophy. 365–373.
- van Inwagen, P. (2000). Free will remains a mystery. *Philosophical Perspectives* 14: 1–19.
- van Inwagen, P. (2017). The problem of fr** w*ll. In: *Thinking about Free Will*, Cambridge: Cambridge University Press. 192–209.
- Vihvelin, K. (2008). Compatibilism, incompatibilism, and impossibilism. In: *Contemporary Debates in Metaphysics* (ed. J. Hawthorne, T. Sider and D. Zimmerman). Malden: Blackwell. 303–318.
- Vihvelin, K. (2013). *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. New York: Oxford University Press.
- Vilhauer, B. (2012). Taking free will skepticism seriously. *Philosophical Quarterly* 62: 833–852.
- Williams, B. and Nagel, T. (1976). Moral luck. *Proceedings of the Aristotelian Society, Supplementary Volumes* 50: 115–135 + 137–151.

Part I

Preliminaries

Logical and Theological Fatalism

ALICIA FINCH

1 Introduction

Fatalism is the thesis that no one acts freely, where:

(FA) Agent *S* freely performs act *A* at time *t* =_{df.} (i) *S* performs *A* at *t* and (ii) for some time *t*- such that *t*- ≠ *t*, it is up to *S* at *t*- whether *S* performs *A* at *t*

and:

(UP) It is up to *S* at *t*- whether *S* performs *A* at *t* =_{df.} *S* at *t*- is both (a) able to perform *A* at *t* and (b) able to refrain from performing *A* at *t*,

so that (FA) is equivalent to:

(FA*) Agent *S* freely performs act *A* at time *t* =_{df.} (i) *S* performs *A* at *t* and (ii) for some time *t*- such that *t*- ≠ *t*, *S* at *t*- is both (a) able to perform *A* at *t* and (b) able to refrain from performing *A* at *t*.

This definition presupposes a *leeway* rather than a *source* theory of free action: according to leeway theorists, free action “requires alternative possibilities”; source theorists, by contrast, contend that if an agent is (in some relevant sense) the *source* of an action, the agent acts freely even if she lacks alternative possibilities.¹ As the ensuing discussion ought to make clear, the debate over fatalism arises only if a leeway theory of free action is presupposed.

Some leeway theorists will balk at a definition of free action according to which the time at which an agent is “able to do otherwise” is necessarily distinct from the time at which the agent acts freely.² In order to address their concern, it will be useful to consider the notion of a *time*. Following Finch and Rea (2008), I acknowledge that times might be thought of either as *abstract* states of affairs³ or as *concrete* events, and I offer that “Abstract times might

fruitfully be thought of as *present-tense maximal state of affairs*.” In defining this notion, they stipulate that:

[A] state of affairs [O] is *future-directed* just in case either [O]’s obtaining entails that some contingent thing will exist or [O]’s obtaining entails that no contingent thing will exist; [a] *past-directed* state of affairs [is defined] in the obviously parallel way. Then a state of affairs [O] is present-tense maximal if and only if, for every atomic state of affairs [O’] that is neither future-directed nor past-directed, either [O] includes [O’] or [O] precludes [O’].⁴(10)

And that:

One state of affairs includes another just in case the obtaining of the first state of affairs entails the obtaining of the second. One state of affairs precludes another just in case the obtaining of the first entails that the second does not obtain. (10, fn. 11)

Finch and Rea add that “A concrete time might then be thought of as the event of some particular abstract state of affairs obtaining” (10). In what follows, I will use the term *time* to refer to present-tense maximal concrete events. However, it should be clear that nothing of substance hinges on using this term in this way;⁵ for the purposes of this essay, what matters is that present-tense maximal concrete events do, in fact, occur.

Indeed, at the moment, what matters is that, given the definition of *S*’s freely performing *A* at *t*, *t* is a present-tense maximal concrete event that includes *S*’s performing *A*. In order to see *why* this should matter, let us consider, first, that it seems uncontroversial to assume that:

(Able) *S* at *t* - is able to perform *A* at *t*.

entails:

(Possible) It is possible⁶ that: (i) *t* - occurs and (ii) *S* performs *A* at *t*.

And, by extension, it seems uncontroversial that if *S* at *t* - is *able to* refrain from performing *A* at *t*, it is (broadly logically) *possible that* (i) *t* - occurs and (iii) *S* refrains from performing *A* at *t*. But given the preceding definition of a time, and given that (*ex hypothesi*) *t* includes *S*’s performing *A*, it follows that it is (broadly logically) *impossible that* (iv) *t* occurs and (iii) *S* refrains from performing *A* at *t*. It follows, in other words, that if *S* at *t* - is both able to perform *A* at *t* and able to refrain from performing *A* at *t*, *t* ≠ *t*.

Having specified the relevant definition of free action, I can move on to distinguishing between logical and theological fatalism. According to *logical fatalism*, the thesis that no one acts freely is entailed by the definition of free action and the logical principle of *bivalence*, which is the thesis that:

(Bivalence) For any proposition *p*, either *p* is true or *p* is false and *p* is not both true and false. (i.e., every proposition has exactly one truth value and there are no truth values other than *truth* and *falsity*.)

The pith of the argument for logical fatalism is this:

Every proposition is either true or false. Suppose that the proposition <Agent *S* performs act *A* at time *t*> is true. If this proposition is *ever* true, it is *always* true. And if this proposition is *always* true, there is never a time at which it is up to *S* whether it is true or false. (Indeed, if this proposition is *always* true, it is true long before *S* comes into existence and, hence, long before *S* is able to do anything at all.) But if there is never a time at which it is up to *S* whether <Agent *S* performs act *A* at time *t*> is true or false, *S* does not freely perform *A* at *t*. Since this point generalizes to any agent, any act, and any time, free action is impossible.

According to *theological fatalism*, the thesis that no one acts freely is entailed by the definition of free action and the existence of an essentially omniscient God, where the definition of *essential omniscience* entails the truth of bivalence:

(EDO) God is essentially omniscient =_{df.} For any proposition p , (i) either God knows that p is true or God knows that p is false, (ii) God knows that p is not both true and false, and (iii) it is impossible⁷ for God to be mistaken about the truth value of any proposition.⁸

Here is the pith of the argument for theological fatalism (which is sometimes called “the problem of freedom and foreknowledge”):

Every proposition is either true or false and God knows the truth value of each proposition. Suppose that the proposition $\langle \text{Agent } S \text{ performs act } A \text{ at time } t \rangle$ is true. If this proposition is true, God has *always known* that it is true. And if God has *always known* that it is true, there is no time at which it is up to S whether or not God knows that it is true. (Indeed, if God has *always known* that it is true, God knows its truth long before S comes into existence and, hence, long before S is able to do anything at all.) But if there is never a time at which it is up to S whether or not God knows that $\langle \text{Agent } S \text{ performs act } A \text{ at time } t \rangle$ is true, S does not freely perform A at t . Since this point generalizes to any agent, any act, and any time, free action is impossible.

In what follows, I will offer formal presentations of these arguments and consider what responses are available to opponents of fatalism. In particular, I will consider responses that involve rejecting (i) the arguments’ validity; (ii) the arguments’ assumptions about whether it could be up to an agent whether a proposition has always been true or whether it could be up to an agent whether God has always known that a true proposition is true; (iii) the arguments’ assumptions about the relationship between truth values and times, on the one hand, or God and times, on the other; and (iv) the arguments’ assumptions about whether propositions can change their truth values or whether an essentially omniscient God can acquire knowledge over time.

2 Formulating the Arguments⁹

In an attempt to present the arguments as clearly and precisely as possible, I will stipulate that:

“ p_A ” designates the proposition that S performs A at t .

Its being up to S at time t - whether p_A is (identical to) its being up to S at time t - whether S performs A at t .

“ $N_{s,t}p$ ” designates: p & it is not up to S at t whether p .

$N_{s,t}p$ is equivalent to (i) p & S at t is unable to render p false and (ii) p & there is nothing S at t can do such that, if S were to do it, p would be false.¹⁰

In addition, I note that I will rely on an inference principle similar to (but different from) van Inwagen’s famed “principle β ” (see, e.g., van Inwagen 1983, 2000):

(Transfer) $\{N_{s,t}p, \Box(p \rightarrow q)\} \vdash N_{s,t}q$

While one might be able to formulate the arguments for fatalism without explicitly appealing to a β -like principle, the strongest versions of the arguments will include such an appeal.

With this, I turn to the premises on which the arguments for logical and theological fatalism rely. I have already mentioned Bivalence and Essential Divine Omniscience. Both arguments also depend on some variation on either the Principle of the Fixity of the Past or the Principle of the Fixity of the Present. The former is the principle that:

(FP): Necessarily, for any agent S , any proposition p , and any time t , if (i) p describes a state of affairs that obtains prior to t , (ii) it is not up to S at or after t whether p

while the latter is the principle that:

(FPr): Necessarily, for any agent S , any proposition p , and any time t , if (i) p describes a state of affairs that obtains at t , (ii) it is not up to S at t whether p .

In what follows, I will present the arguments in terms of the Fixity of the Past. I will do so because this seems to be standard practice, probably because it seems more dialectically effective than the alternative. I note, though, that (i) given the definition of free action, the Fixity of the Present is trivially true and (ii) nothing of philosophical significance hinges on my presenting the arguments in terms of the Fixity of the Past rather than the Fixity of the Present.

Having stated what the two arguments for fatalism have in common, I turn my attention to the argument for logical fatalism. It depends on the truth of these principles:¹¹

(Truth-at- t): Necessarily, for any proposition p , if p is true, there is some time t such that p is true at t .

And:

(Immutability): Necessarily, for any proposition p , for any time t , and for any time t^* , p is true at t if and only if p is true at t^* .

The first says that each true proposition is true *at a time*; the second says that propositions do not change their truth values *across times*. I will consider these principles in more detail when I discuss attempts to reject them. For now, I will simply acknowledge that while it is possible to formulate the argument without explicitly invoking these principles, the argument succeeds only if these (or relevantly similar) principles are true.

With the stipulations that (i) " t_{-1B} " designates a time approximately one billion years before time t -, (ii) S exists at t -, and (iii) S did not come into existence until long after t_{-1B} ,¹² the argument for logical fatalism may be presented as:

1. p_A	Assumption
2. $\Box(p_A \leftrightarrow p_A \text{ is true at } t)$	Truth-at- t
3. $\Box(p_A \text{ is true at } t \leftrightarrow p_A \text{ is true at } t_{-1B})$	Immutability
4. $\Box(p_A \leftrightarrow p_A \text{ is true at } t_{-1B})$	2, 3
5. $N_{s,t-}(p_A \text{ is true at } t_{-1B})$	Fixity of the Past
6. $N_{s,t-} p_A$	4, 5 Transfer

I note that while this argument depends on the assumption that p_A is true, the same argument could be made, *mutadis mutandis*, on the assumption that p_A is false. What matters is that, given Bivalence, p_A has one truth value or the other (and not both).

With this, I will offer a formal version of the argument for theological fatalism. In addition to depending on Essential Divine Omniscience and the Fixity of the Past (or Present), this version of the argument depends on the thesis that God is everlasting, where:

(Divine Everlastingness) God is everlasting = *df.* Time t obtains if God exists at t ,

and the principle that the knowledge of God is immutable:

(Immutable Knowledge) Necessarily, for any time t , for any time t^* , and for any proposition p , God at t knows p if and only if God at t^* knows p .¹³

This principle of Immutable Knowledge follows from Immutability, Divine Everlastingness, and Essential Divine Omniscience. Here, then, is the argument for theological fatalism:

1' . p_A and God exists	Assumption
2' . $\Box(p_A \leftrightarrow \text{God at } t \text{ knows } p_A)$	1', EDO, Divine Everlastingness
3' . $\Box(\text{God at } t \text{ knows } p_A \leftrightarrow \text{God at } t_{1B} \text{ knows } p_A)$	Immutable Knowledge
4' . $\Box(p_A \leftrightarrow \text{God at } t_{1B} \text{ knows } p_A)$	2', 3'
5' . $N_{s,t-}(\text{God at } t_{1B} \text{ knows } p_A)$	Fixity of the Past
6' . $N_{s,t-} p_A$	4', 5' Transfer

In the next section, I will begin to consider different responses to the two arguments.

First, though, I will pause to note that although some philosophers (e.g., Zagzebski) find the argument for theological fatalism more compelling than the argument for logical fatalism, Ted A. Warfield (1997) has pointed out that if an essentially omniscient God exists necessarily, then for any proposition p , $\langle p \text{ is true at } t \rangle$ is logically equivalent to $\langle \text{God at } t \text{ knows } p \rangle$. Warfield further offers that:

If p and q are logically consistent, then p is consistent with any proposition that is logically equivalent to q .

Warfield's point is that if logical fatalism is false, some proposition p_A is such that (i) $\langle p_A \text{ is true at } t_{1B} \rangle$ and (ii) $\langle p_A \text{ is true at } t_{1B} \rangle$ is logically consistent with $\langle \text{It is up to some agent } S \text{ at some time } t \text{ - whether } p_A \rangle$. But given the aforementioned principle, and given that $\langle p_A \text{ is true at } t_{1B} \rangle$ is logically equivalent to $\langle \text{God at } t_{1B} \text{ knows } p_A \rangle$, it follows that if logical fatalism is false, some proposition p_A is such that (i) $\langle \text{God at } t_{1B} \text{ knows } p_A \rangle$ and (ii) $\langle \text{God at } t_{1B} \text{ knows } p_A \rangle$ is logically consistent with $\langle \text{It is up to some agent } S \text{ at some time } t \text{ - whether } p_A \rangle$. In other words, it follows that if logical fatalism is false, so is theological fatalism (and vice versa).

I will not dwell on assessing Warfield's argument. In what follows, though, I will emphasize that for every response to the argument for logical fatalism, there is an analogous response to the argument for theological fatalism (and vice versa).

3 Response 1: The Arguments are Invalid

The most elegant response to both arguments is to simply deny the validity of the Transfer Principle (and other β -like principles).¹⁴ On behalf of this response, one might note that the strongest case for β -like principles seems to consist of pointing to them and asking "How could they *not* be valid? Doesn't it seem obvious that they are?" Opponents of fatalism could suggest that, actually, the validity of β -like principles is not at all obvious and that unless someone presents them with an argument for the principles' validity, they will continue to reject the fatalists' conclusion.

This response confronts at least two problems. First, β -like principles *do* strike many participants in the debate as valid. Indeed, some defenders of these principles find the core insight so compelling that they simply will not abandon the principles, even if they concede that a *particular* β -like principle must be jettisoned in favor of a reformulation.¹⁵ Second, many participants in the free will debate take it for granted that (i) the Consequence Argument for the incompatibility of free action and determinism is valid only if some β -like principle is valid, and that (ii) without the Consequence Argument, the case for the incompatibility of free action and determinism is relatively weak. Opponents of fatalism who are *libertarians* (who, that is, are incompatibilists about free action and determinism while affirming the thesis that some agents do, in fact, act freely) might reasonably conclude that rejecting the validity of β -like principles is too high a price to pay for a response to fatalism. Fortunately for them, the other responses to fatalism that I consider are consistent with libertarianism.

4 Response 2: Ockhamism

In this section, I will consider Ockhamism, a response to fatalism that challenges the Principle of the Fixity of the Past (and Present). I will follow the standard practice of presenting Ockhamists as drawing a distinction between so-called *hard facts* and *soft facts*. While there is no consensus on how to draw the distinction in question, the basic idea is that “A hard fact about the past is entirely about the past whereas a soft fact is not: a hard fact about, say, t_{1B} is a fact whose obtaining is entirely independent of whatever might happen after t_{1B} , whereas a soft fact about t_{1B} somehow depends on, or involves, or includes events that take place at later times”¹⁶ (Finch and Rea p. 3). Given the definitions already introduced, we can say that a hard fact is included in a time (that is, a present-tense maximal concrete event) that has occurred and a soft fact is a future-directed state of affairs. Of course, whether a fact is hard or soft is relative to a time.

The Ockhamist response to both logical fatalism and theological fatalism may be construed as a rejection of the Principle of the Fixity of the Past (or Present). More precisely, it amounts to the position that the Principle of the Fixity of the Past is ambiguous between the Principle of the Fixity of the Hard Past:

(FHP) Necessarily, for any agent S , any proposition p , and any time t , (i) if p describes a state of affairs O that is a hard fact at t , (ii) it is not up to S at or after t whether p (is true),

and the Principle of the Fixity of the Soft Past:

(FSP) Necessarily, for any agent S , any proposition p , and any time t , (i) if p describes a state of affairs O that is a soft fact at t , (ii) it is not up to S at or after t whether p (is true).

While the former is true, Ockhamists say, the latter – the very principle on which the arguments for fatalism depend – is false.

In order to appreciate why the Ockhamists reject the Principle of the Fixity of the Soft Past, let us consider a specific example. In particular, let us suppose that p_M is true where:

“ p_M ” designates the proposition that Mary marries Harry at t_M .

and:

“ t_M ” designates noon on March 13, 3013;

Given that the current year is 2021, it is *now* a soft fact that p_M is true.

Ockhamists will insist that the soft fact that Mary marries Harry at t_M is consistent with the existence of a time t_{-M} such that (i) t_{-M} is earlier than t_M and (ii) Mary at t_{-M} is able to do something such that, if she were to do it, p_M *would* be false. Perhaps Mary at t_{-M} is able to cancel the wedding, convince Harry that they should elope before March 13, 3013, or shout, “No!” when asked whether she takes Harry as her lawfully wedded spouse. To be clear, Ockhamists do not suggest that Mary is able to do something that would *change* the past. Rather, they contend that the truth of p_M is consistent with Mary’s being able to do something such that, if she *were* to do it, p_M *never would have been true* in the first place. Given that Mary and Harry are, in fact, married at t_M , p_M has always been true and God has always known p_M .

Ockhamists emphasize that p_M has always been true *because of* what happens at t_M , and not the other way around. The order of dependence (or priority) is crucial: the soft facts depend on hard facts in a way that the hard facts do not depend on the soft. Unfortunately, Ockhamists do not typically explain how exactly to read “because of” or what sort of dependence or priority they have in mind. Finch and Rea (2008) have suggested, though, that if *eternalism* about the metaphysics of time is correct, Ockhamism succeeds even without such an explanation. They offer a standard construal of eternalism, according to which eternalism is “the thesis that past, present, and future objects (and, by extension, events) exist” (10).¹⁷ Given the assumption that an event exists if and only if it occurs, eternalism is true only if all past, present, and future events occur. Of course, this is not to say that all past, present, and future events occur *at the same time* (which would be absurd); rather, all the events that occur stand in relations of *earlier-than*, *simultaneous with*, and *later-than* to one another. Given this definition, eternalism implies that if agent S performs act A at time t , S ’s performance of A at t occurs; that is, S ’s performance of A at t exists in the concrete world. And, as Finch and Rea indicate:

[W]e are fully prepared, in the ordinary case, to think that the proposition that S performs A is ontologically dependent on S ’s performance of A and, moreover, that S ’s performance of A is ontologically prior to the truth of the proposition that S performs A . The eternalist Ockhamist’s point is that, however we ordinarily understand the relationship between true propositions about agents’ action and the agents’ actions themselves, this is how we should understand the relationship between true propositions like [It was true at t_{-1B} that S performs A at t] and S ’s performance of A at t . (11–12)

If they are right, then if eternalism is true, there is a straightforward sense in which its being true at t_{-1B} that S performs A at t is dependent on (the concrete event that is) S ’s performing A at t . Moreover, they might have added, there is also a straightforward sense in which God’s knowing at t_{-1B} that S performs A at t depends on (the concrete event that is) S ’s performing A at t . Eternalist Ockhamists need not provide a full account of knowledge in general or divine knowledge in particular; they need only point out that there is nothing especially exotic about the suggestion that God’s knowledge of what happens at t depends on what happens at t , but not vice versa.

Of course, Finch and Rea could have made the same point about Ockhamism and ontological dependence without invoking eternalism per se: any account of the metaphysics of time according to which all future objects and events determinately exist would do just as well. In the final section of this essay, I will consider which dialectical options are available to opponents of fatalism who reject the thesis that the future determinately exists. First, though, I will turn my attention to another strategy altogether.

5 Response 3: Propositions Are Not True *at Times*; God Does Not Have Knowledge *at Times*

In this section, I will consider responding to the argument for logical fatalism by rejecting Truth-at-*t* and, analogously, responding to the argument for theological fatalism by rejecting Divine Everlastingness. The former response challenges the very idea that propositions are true *at times*; the latter response challenges the claim that God exists *at times*. I note that while these responses are analogous to one another, they do not seem to stand or fall together: it seems that one could accept Truth-at-*t* while rejecting Divine Everlastingness, or vice versa.

While Truth-at-*t* might seem innocuous at first, Peter van Inwagen raises a significant challenge against it. In particular, he suggests that it is simply *nonsense* to assert that a proposition is true at a time. He makes this suggestion by considering similar expressions (e.g., “true at some particular moment,” “true at every moment,” “became true,” “remained true,” “is unchangeably true”) and contending that he cannot “see what these phrases mean if they are used as they are used in the above argument for fatalism” (35). He concedes that if someone were to say, “Municipal bonds are a good investment,” and if someone else were to reply, “That used to be true but it isn’t true anymore,” his respondent’s words “would be a model of lucidity” (35). But these words are lucid precisely because his respondent could express the same thoughts without resorting to talk of propositions true at times. For instance, he could reply, “While municipals bonds used to have a high rate of return, they do not have a high rate of return today.” How exactly could one capture the meaning of “ p_A was true 1 billion years ago” without using the notion of truth at a time? Van Inwagen evaluates several proposals for rephrasing and argues that each is meaningless. As such, he concludes that the argument for logical fatalism rests on a faulty assumption about the relationship between truths and times.

The analogous response to the argument for theological fatalism can be traced back to Boethius, a sixth-century Christian philosopher. Boethius considered God the one concrete object who exists *atemporally* or *eternally* (that is, outside of the temporal order). According to the thesis of Divine Eternity:

(Divine Eternity) God exists and for any time t , God does not exist at t .

Of course, if God does not *exist* at any time, God does not *have knowledge* at any time. As such, it is simply false that God at t_{-1B} knows the truth value of p_A . While the principle of the Fixity of the Past may be true, its truth has nothing to do with God’s knowledge.

But while Boethius thought that God neither exists nor has knowledge at times, he certainly did not think God was ignorant of what happens within the temporal order. As Linda Zagzebski explains:

The way Boethius describes God’s cognitive grasp of temporal reality, all temporal events are before the mind of God at once. To say “at once” or “simultaneously” is to use a temporal metaphor, but Boethius is clear that it does not make sense to think of the whole of temporal reality as being before God’s mind in a single *temporal* present. It is an *atemporal* present, a single complete grasp of all events in the entire span of time. (2016)

This explanation suggests that Boethius endorsed eternalism with respect to the metaphysics of time. Insofar as God is eternal, none of the events included in the temporal order is temporally closer to God than any other; insofar as God is omniscient, God knows exactly which

events occur, how these events are temporally ordered with respect to one another, and how objects and events that exist at different times are diachronically related.

Objections to Boethianism abound.¹⁸ For instance, philosophical objections challenge the very coherence of the position while theological objections question whether Boethius's picture of God is consistent with other theses about the nature of God and God's relationship to creation. In the present context, though, the most pressing objection is one that has been raised by Zagzebski: even if Boethius is correct about the relationship between God and the temporal order, the threat of theological fatalism remains. After all, as Zagzebski notes, "we have no more reason to think we can do anything about God's timeless knowing than about God's past knowing. The timeless realm is as much out of our reach as the past." Indeed, it seems that the theological fatalist may simply reformulate the argument, replacing the principle of the Fixity of the Past (or Present) with a principle of the Fixity of the Eternal:

(FE): Necessarily, for any agent S , any proposition p , and any time t , if (i) p describes a state of affairs that (a) obtains and (b) does not obtain at any time, then (ii) it is not up to S at t whether p .

Stipulating that:

" p_{GEKA} " designates the proposition that God eternally knows p_A ,

the Boethian analogue of the argument for theological fatalism may be formulated such that:

1" . p_A	Assumption
2" . $\Box(p_A \leftrightarrow p_{GEKA})$	EDO, Divine Eternity
3" . $N_{s,t} p_{GEKA}$	Fixity of the Eternal
4" . $N_{s,t} p_A$	2", 3" Transfer

So, merely asserting that God is eternal does not undermine the argument for theological fatalism.

But there seems to be more to the Boethian solution than this mere assertion. If we construe Boethianism as the conjunction of the theses that (i) God is eternal and (ii) standard eternalism is true, and if Boethians make the reasonable assumption that God's eternal knowledge of what happens at t depends on what happens at t , but not vice versa, the Boethian response to theological fatalism seems importantly similar to the Ockhamists' (though without the distinction between hard and soft facts). In order to see that this is so, let us return to p_M , the proposition that Mary marries Harry at t_M . Boethians can offer that although God eternally knows p_M , this eternal knowledge is consistent with the claim that there is a time t_{-M} such that (i) t_{-M} is earlier than t_M and (ii) Mary at t_{-M} is able to do something such that, if she were to do it, it would be false that God eternally knows p_M ; in this case, God would eternally know that p_M is false. The Principle of the Fixity of the Eternal is false, Boethian might say, because what God eternally knows depends on what happens in the temporal order, and not vice versa. Indeed, if Mary would have decided to postpone the wedding, God would have known from eternity that p_M was false. But given that Mary did, in fact, go through with the wedding, God eternally knows that she does so.

Of course, this response leaves one with many questions about the relationship between an atemporal God and a temporal concrete order. Unfortunately, it is beyond the scope of this essay to explore these questions, just as it is beyond the scope of this essay to explore other objections to Boethianism.

Before moving on, though, it seems crucial to raise one further question about the Fixity of the Eternal: does this principle undermine van Inwagen's response to the argument for logical fatalism? After all, if propositions are not true *at times*, and if a proposition p is true, it seems that the state of affairs of p 's *being true* obtains eternally (or "outside the temporal order," or "in the timeless realm," to use Zagzebski's phrase).

Though I do not presume to know what van Inwagen himself would say about this objection, at least two responses seem available. First, one might object to the thesis that propositions are "eternally true" just as van Inwagen objects to the thesis that propositions are "true at times." In adopting the analogous strategy, one might ask what is meant by " p is eternally true." Taking a cue from van Inwagen, one might point out that this is a metaphysician's turn of phrase if ever there was one, and wonder why we should suppose that it is meaningful. One might add that propositions are either true *simpliciter* or not at all: they are neither true at times nor true eternally.

Then again, one might also adopt the strategy I have recommended to the Boethian: affirm standard eternalism about the metaphysics of time and insist that the Principle of the Fixity of the Eternal is false because what is eternally true depends on what happens in the temporal order, and not vice versa. As far as I know, van Inwagen has never committed himself to the truth of eternalism, let alone the Boethian-style response to the principle of the Fixity of the Eternal. Nevertheless, it seems worthwhile to note this dialectical option.

Further discussion of this issue is beyond the scope of this essay and, as such, I turn to a final type of response to the arguments for fatalism.

6 Response 4: The Open Future View and Open Theism

In this section, I will consider both the response to logical fatalism offered by the *open future view* and the response to theological fatalism offered by *open theism*. If the open future view is correct and God is temporally located, open theism is true; however, the converse does not hold: van Inwagen, for instance, is an open theist who affirms that God is temporally located but denies the open future view. In what follows, I hope to make the dialectical options clear.

With respect to the open future view, Amy Seymour (manuscript) offers that:

According to [the open future view], if the future is to be open, the future must be metaphysically unsettled ... This unsettledness is not merely epistemic or linguistic. It is not that we merely do not know what the future holds or that our terms cannot precisely capture what will happen – something about the nature of reality itself is unsettled. (1)

Given this definition, the open future view is inconsistent with standard eternalism and any other account of the metaphysics of time that entails that the future determinately exists.

According to the open future view, if p_+ is a contingent proposition that purports to describe a state of affairs that obtains in the future and does not obtain *now*, p_+ is not true *now*, but it *might become true*. If, for instance, Mary does indeed marry Harry at noon on March 13, 3013, the proposition that Mary marries Harry at noon on March 13, 3013 will become true at the relevant future time. Since this proposition is not true now but will become true, its truth value will change. In short, Immutability is false and, hence, so is the relevant premise of the argument for logical fatalism:

3. $\Box(p_A \text{ is true at } t \leftrightarrow p_A \text{ is true at } t_{1B})$.

This response is at least *prima facie* plausible: after all, it amounts to the claim that if a state of affairs does not yet obtain, it is not yet true that it will obtain.

In terms of working out the details, proponents of the open future view have two dialectical options: (i) denying bivalence; (ii) insisting on the falsity of all future contingent propositions that are not entailed by propositions that are true now (hereafter, *future contingents*). I will consider each view in turn.

According to open futurists who deny bivalence, future contingents are neither true nor false: while some bivalence deniers contend that such propositions have no truth value at all, others insist that they have a truth value *other than* truth or falsity. Those in the latter group typically embrace *multivalent* logic systems, e.g., the three-valued logic systems offered by Jan Kukasiewicz and Stephen C. Kleene. As Theodore Sider explains, “The third truth value is (in most cases, anyway) supposed to represent sentences that ... have some other status. This other status could be taken in various ways, depending on the intended application, for example: ‘meaningless’, ‘undefined’, or ‘indeterminate’ ” (Sider 2010: 73).

At first blush, it may seem reasonable to suggest that propositions about the future are indeterminate in truth value: after all, since the future has not yet obtained, one might say, it is indeterminate what will (or will not) happen. One must consider, though, that classical logic is bivalent, so one cannot consistently deny bivalence without admitting that “classical logic is wrong – that it provides an inadequate model of (genuine) logical truth and logical consequence” (Sider 72). The rejection of classical logic may seem to be too high a price to pay. Then again, as I have said, various multivalent logics have been developed. Moreover, one might be inclined to accept a multivalent logic because it can help not only with future contingents but also with sentences that (i) involve vague terms, (ii) express propositions with false presuppositions (e.g., that the king of France is bald), or (iii) seem to include references to fictional entities. Given the resources afforded by multivalent logics, some proponents of the open future view conclude that the rejection of classical logic is a reasonable price to pay.

Other proponents of the open future view, by contrast, maintain their commitment to classical logic and endorse the position known as “all falsism” (so named by Amy Seymour, manuscript) and “Russellian open futurism” (so named by Patrick Todd 2016). They agree with bivalence deniers that future contingents are *not true*; indeed, they insist that a proposition *p* is true if and only if either the corresponding state of affairs obtains or *p* is entailed by propositions that correspond to states of affairs that obtain. Since future contingents correspond to states of affairs that do not yet obtain and are not entailed by propositions of the relevant sort, it obviously follows that these propositions are not true. Here is where their disagreement with bivalence deniers becomes salient: all-false theorists insist that, necessarily, a proposition *p* is not true if and only if *p* is false; a proposition’s not being true is both necessary and sufficient for its falsity. Since future contingents correspond to states of affairs that do not yet obtain, they are not yet true and, hence, they are *all false*.

With this, one might object that this position entails the falsity of the Law of Non-Contradiction (according to which it is necessarily the case that $\neg(p \ \& \ \neg p)$). All-false theorists will insist that this objection confuses propositions of the form <It is false that *O* will obtain> with propositions of the form <*O* will not obtain>; the former but not the latter is the negation of <*O* will obtain>; the former but not the latter is consistent with the truth of <It is false that *O* will not obtain>. To return to the case of Mary and Harry: according to the all-false theorist, <It is false that Mary will marry Harry at t_M > and <It is false that Mary will not marry Harry at t_M >; the former is the negation of <Mary will marry Harry at t_M >; and the negation of <Mary will marry Harry at t_M > is not equivalent to <Mary will not marry Harry at t_M >.

While one might balk at the all-false theorists' claim that false propositions become true, one should not confuse this claim with a violation of the Principle of Non-Contradiction.

Whether open futurists go the route of bivalence denial or all-falsism, they must admit that their view *seems* at odds with various practices of ordinary life. After all, we constantly form beliefs and make statements about the future and, in so doing, we *seem* to proceed on the assumption that these propositions are true. The open futurist must either concede that we are constantly mistaken about the nature of reality or explain our behavior in such a way that it does not, despite appearances, depend on false assumptions about the future. Unfortunately, it is beyond the scope of this essay to offer an extensive discussion of either variation on the open future view.

Instead of dwelling on how well the open future view fares as a response to the argument for logical fatalism, I will shift my focus to *open theism*, the analogous response to the argument for theological fatalism. According to open theists, the thesis of Immutable Knowledge is false and, as such, so is the premise that:

3'. $\Box(\text{God at } t \text{ knows } p_A \leftrightarrow \text{God at } t_{1B} \text{ knows } p_A)$

While God at t knows that S performs A at t , the open theist denies that God at t_{1B} knows that this is case: God's knowledge increases over time.

Open theists do not take themselves to be denying the essential omniscience of God and, in fact, there are three dialectical options for open theists who seek to maintain their commitment to this theological principle.¹⁹ First, open theists might embrace the open future view along with "all-falsism"; in this case, they might define essential divine omniscience such that:

(EDO') God is essentially omniscient $\equiv_{\text{df}}$ For any proposition p , (i) either God at t knows that p is true or God at t knows that p is false, (ii) God at t knows that p is not both true and false, and (iii) it is impossible for God to be mistaken about the truth value of any proposition

In this case, open theists would contend that since it is false at t_{1B} that S performs A at t , God at t_{1B} knows that it is false that S performs A at t . When, at t , it becomes true that S performs A at t , God at t will come to know what S does at t .

Second, open theists might adopt the open future view while rejecting bivalence. They would offer something like this as a definition of essential divine omniscience:

(EDO'') God is essentially omniscient $\equiv_{\text{df}}$ For any proposition p , (i) God at t knows the truth value of p at t and (ii) it is impossible for God to be mistaken about the truth value of any proposition.

In this case, they will insist that since it is neither true nor false at t_{1B} that S performs A at t , God at t_{1B} knows that it is neither true nor false that S performs A at t . When, at t , it becomes true that S performs A at t , God at t comes to know that S does so.

But there is a third dialectical option for the open theist, as discussed by Peter van Inwagen.²⁰ He presents his discussion of theological fatalism in the context of discussing the problem of evil (see his (2006)), initially offering that:

A being is omniscient if, for every proposition, that being believes either that proposition or its denial, and it is metaphysically impossible for that being to have false beliefs. (26)

But then he asks:

Why not say that even an omniscient being is unable to know certain things – those such that its knowing them would be an intrinsically impossible state of affairs? Or we might say this: an omnipotent being is also omniscient if it knows everything it is able to know. (82).

I take it that van Inwagen is suggesting that:

(EDO'') God is essentially omniscient =_{df.} For any proposition p such that it is possible for God at t to know p or its denial, God at t believes either p or its denial, and it is impossible for God to be mistaken about the truth value of any proposition.

And that he is further suggesting that:

For any proposition p_A such that p_A is a proposition about a free act that agent S performs at time t , and for any time t^- such that t^- is earlier than t , it is impossible for God at t^- to know whether p_A or its denial is true.

On van Inwagen's picture, the complete state of the world prior to t fails to determine whether S performs A at t . As such, God withholds belief, prior to t , about S 's performance of A at t : Until t is present, God simply cannot know whether S 's performance of A at t obtains.

In response, one might object that this definition seems strained, as if the only reason to endorse it is to escape the problem of theological fatalism. One might demand a reason to think that there is a difference between (i) true propositions and (ii) true propositions such that it is possible for an essentially omniscient and everlasting God to know that they are true. Unfortunately, it is unlikely that we will find such a reason without embarking on a thorough discussion of divine knowledge, which is beyond the scope of this essay.

With respect to the first two varieties of open theism (those corresponding to all-falsism and bivalence denial), theological objections seem more prevalent than logical ones. Indeed, much ink has been spilled in considering whether open theism is consistent with orthodox Christianity. Many of its defenders consider themselves orthodox Christians (van Inwagen and Hasker, e.g.), but some opponents suggest that open theists ought to be regarded as heretics. While open theism is clearly inconsistent with, for instance, Roman Catholicism, its compatibility with other Christian traditions is not so obvious. The theological disputes over open theism certainly cannot be settled here.

7 Conclusion

The arguments for logical and theological fatalism are structurally similar, with both depending on (i) the definition of free action, (ii) some variation on the Transfer Principle, and (iii) the principle of the Fixity of the Past (or Present). Moreover, while the argument for logical fatalism depends on (iv) Bivalence, (v) the thesis that propositions are true at times, and (vi) the thesis that propositions do not change their truth values, the argument for theological fatalism depends on the analogous theses that (iv') God is essentially omniscient, (v') God exists at every time, and (vi') God's knowledge is immutable.

As I suggested, the most elegant response to both arguments is the rejection of the Transfer Principle. However, some participants in the debate find the Principle so obvious that they will find this strategy prohibitively costly; moreover, as I explained, incompatibilists regarding free action and determinism should be loath to abandon this principle, given that it is crucial for the Consequence Argument for incompatibilism.

Ockhamism is available as a response to both logical and theological fatalism: Ockhamists distinguish between the Fixity of the Hard Past and the Fixity of the Soft Past, and insist that while the former is true, the latter is false; given that the arguments for fatalism depend upon the latter, Ockhamists contend that they have quelled the fatalists' threat to free action. Following Finch and Rea, I suggested that Ockhamists ought to embrace standard eternalism with respect to the metaphysics of time.

I made a similar suggestion with respect to the Boethian response to the argument for theological fatalism. According to this response, the argument fails because God does not have knowledge *at times*. Following Zagzebski, I suggested that the argument could be reformulated in terms of the Fixity of the Eternal rather than the Fixity of the Past (or Present). I then pointed out that if eternalism about the metaphysics of time is true, and if God's eternal knowledge depends on which concrete events actually occur, the Boethian can make a case against the Fixity of the Eternal analogous to the Ockhamist's case against the Fixity of the Soft Past. Unfortunately, it is beyond the scope of this essay to delve more deeply into either Ockhamism or Boethianism.

Alongside the Boethian response to theological fatalism, I considered van Inwagen's response to the argument for logical fatalism: according to van Inwagen, the argument fails insofar as it includes the premise that propositions are true at times – a proposition that van Inwagen takes to be meaningless.

Finally, I considered the open future view and open theism. According to open future views, the future is unsettled and, as such, some propositions about the future change truth values. While some open theists also adhere to open future views, van Inwagen is an exception: though he thinks that all propositions have true values, he thinks that propositions about agents' future free acts are unknowable; he offers a definition of Essential Divine Omniscience that allows him to say that God is essentially omniscient even though God lacks knowledge of the relevant truth values. As one would expect, philosophers' theological commitments play a crucial role in their evaluations of open theism.

There is no obvious response to either fatalist argument. In formulating a response, one must consider not only whether one is a compatibilist about free will and determinism but also one's views on the metaphysics of time and the truth values of propositions (and, when considering the argument for theological fatalism, one's theological commitments). Indeed, debates over logical and theological fatalism are strongly connected to other philosophical debates, and acknowledging these connections seems crucial for moving the debates forward.

Notes

- 1 For discussion of the debate between leeway and source theorists, see Timpe 2017.
- 2 See, e.g., Campbell (2007, 2008, 2010).
- 3 In the present context, we need not be very careful about defining states of affairs: it is enough to say that (i) states of affairs are *ways things are*, (ii) for each state of affairs *O*, there is a corresponding proposition *p*, and (iii) a proposition *p* is true if and only if the corresponding state of affairs *O* obtains.
- 4 Anyone concerned about relativity theory can add the qualifier "from a frame of reference."
- 5 Moreover, in the present context, nothing of substance hinges on whether relationism or substantivalism with respect to time is true.
- 6 Throughout this essay, "possible" should be read as "broadly logically possible" and "necessary" should be read as "broadly logically necessary."

- 7 Throughout this essay, “impossible” should be read as “broadly logically impossible.”
- 8 Although philosophers disagree about how to define “omniscience” in general and “essential omniscience” in particular, this definition is satisfactory in the present context. For discussion of the notion of omniscience, see Weirenga 2017.
- 9 My formulation of the arguments here tracks my formulation of the argument for logical fatalism in Finch 2017. Moreover, my formulation of the arguments is obviously inspired by Peter van Inwagen’s formulation of the so-called Consequence Argument for the incompatibility of free action and determinism (where determinism is the thesis that, at any given time, there is only one nomologically possible future.) See, e.g., van Inwagen 1983, 2000.
- 10 In the present context, I could just as well have used “might” in place of “would.” As McKay and Johnson 1996 and van Inwagen 2000 make clear, there are certain contexts in which the difference between “would” and “might” conditionals is significant; this context, however, is not one of them.
- 11 More accurately: it depends on the truth of principles *relevantly similar* to these.
- 12 For the purposes of this essay, in order to streamline discussion, I am setting aside the possibility of the sort of time travel familiar from science fiction.
- 13 More accurately: it depends on the truth of principles *relevantly similar* to these.
- 14 For discussion of the validity of β -like principles see Flint 1987.
- 15 See, again, McKay and Johnson 1996 and van Inwagen 2000.
- 16 For discussion of the distinction between hard and soft facts, see Todd 2013.
- 17 To clarify: Finch and Rea assume that if eternalism is true, all past, present, and future objects (and, by extension, events) *determinately* exist and all past, present, and future events *determinately* occur. This is the standard construal of eternalism. By contrast, Elizabeth Barnes and Ross Cameron (2011) present a version of eternalism according to which future objects exist *indeterminately*, so that even though future objects exist, it is “unsettled” what will happen in the future. Their view is a species of *open futurism*, which I will discuss in the final section of this paper.
- 18 For discussion of the debate over Boethianism, see Helm 2014.
- 19 See, e.g., Rhoda 2008 and Tuggy 2007.
- 20 This strategy is also endorsed by William Hasker (1998, 2004).

Bibliography

- Barnes, E. and Cameron, R. (2011). Back to the open future. *Philosophical Perspectives* 25: 1–26.
- Campbell, J. (2007). Free will and the necessity of the past. *Analysis* 67: 105–111.
- Campbell, J. (2008). Reply to Brueckner. *Analysis* 68: 264–269.
- Campbell, J. (2010). Incompatibilism and fatalism: reply to loss. *Analysis* 70: 71–76.
- Finch, A. (2017). Logical fatalism. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), 191–202. New York: Routledge.
- Finch, A. and Rea, M. (2008). Presentism and Ockham’s way out. In: *Oxford Studies in Philosophy of Religion, Volume 1* (ed. J. Kvanvig), 1–17. New York: Oxford University Press.
- Flint, T. (1987). Compatibilism and the argument from unavoidability. *Journal of Philosophy* 84: 423–440.
- Hasker, W. (1998). *God, Time and Knowledge*. Ithaca, NY: Cornell University Press.
- Hasker, W. (2004). *Providence, Evil, and the Openness of God*. New York: Routledge.
- Helm, P. (2014). Eternity. *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), (ed. E.N. Zalta). <http://plato.stanford.edu/archives/spr2014/entries/eternity>. (Accessed on 29th August 2016).
- McKay, T. and Johnson, D. (1996). A reconsideration of an argument against compatibilism. *Philosophical Topics* 24: 113–122.
- Rhoda, A. (2008). Generic open theism and some varieties thereof. *Religious Studies* 44: 225–234.
- Seymour, A. manuscript. All-falsism considered.

- Sider, T. (2010). *Logic for Philosophy*. Oxford: Oxford University Press.
- Timpe, K. (2017). Leeway vs. sourcehood conceptions of free will. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), 213–224. New York: Routledge.
- Todd, P. (2013). Soft facts and ontological dependence. *Philosophical Studies* 164: 829–844.
- Todd, P. (2016). Future contingents are all false! on behalf of a Russellian open future. *Mind* 125: 775–798.
- Tuggy, D. (2007). Three roads to open theism. *Faith and Philosophy* 24: 28–51.
- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.
- van Inwagen, P. (2000). Free will remains a mystery. *Philosophical Perspectives* 14: 1–20.
- van Inwagen, P. (2006). *The Problem of Evil*. Oxford: Clarendon Press.
- Warfield, T. (1997). Divine foreknowledge and human freedom are compatible. *Noûs* 31: 80–86.
- Weirenga, E. (2017). Omniscience. *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), (ed. E.N. Zalta). <https://plato.stanford.edu/archives/spr2017/entries/omniscience>. (Accessed on 27th June 2017).
- Zagzebski, L. 2016. Foreknowledge and free will. *The Stanford Encyclopedia of Philosophy* (Summer 2016 Edition), (ed. E.N. Zalta). <http://plato.stanford.edu/archives/sum2016/entries/free-will-foreknowledge>. (Accessed on 13th June 2016).

Causal Determinism

V. ALAN WHITE

1 Preliminaries

The free will problem has long been a staple of the philosophy curriculum, and students early on learn the terms “compatibilism” and “incompatibilism” as a useful way to sort concepts of free will into mutually exclusive categories. These terms are a pair representative of the strategy of inter-definition by negation, whereby one term is negatively defined off the other, which is taken as basic. Thus for example all possible colors can be sorted into those that are red and not-red. Compatibilism and incompatibilism similarly sort all possible concepts of free will into two categories logically based on some common essential core of definition. Following this terminology, “compatibilism” would appear to be the basic term, with the “in” of “incompatibilism” signifying that it is negatively inter-defined off it, as respectively “red” is basic for defining “non-red.” Taken this way, compatibilism states that *free will, properly defined, is logically compatible with the truth of determinism*. Incompatibilism then states by negation that *it is false that free will, properly defined, is logically compatible with the truth of determinism*.¹

This classic way of sorting concepts of free will into two categories is focused on the common core of whether such a concept logically conflicts with determinism – *incompatibilism* – or does not logically conflict with determinism – *compatibilism*.² What is clear from this is that if one wishes to understand what “free will” means in either case depends on what *determinism* is, since the matter of logical conflict with it is taken to be foundational to structuring the free will problem this way.

There is a strong sense therefore that this issue of determinism has been so pivotal to the free will problem that it guides the very ways we should talk and think about “free will.” I intend to examine how and why this is the case, and to offer an alternative form of determinism to serve as the basis for the incompatibilist/compatibilist distinction.

1.1 Causal Determinism as the Relevant Concept

The first task is to specify what determinism is as properly standing behind the incompatibilist/compatibilist distinction. This is necessary because the term “determinism” is deemed to describe radically different positions distinguishable by affixing qualifiers signaling something about the nature of those differences. Thus, there are forms of *logical* determinism, *theological* determinism, *sociological* determinism, *causal* determinism, and more, and even more specific sub-types within these just named. The last named type – *causal determinism* (and its sub-types) – will be the one needed here, but that claim requires justification.

Clues that single out causal determinism as what is needed here can be mined out of an analysis of the motivation behind forming the incompatibilist/compatibilist distinction in the first place. To do this, we need to see why any concept of free will might or might not conflict logically with some concept of determinism.³ To do that analysis, we need to capture some rock-bottom requirements the use of those two terms – “free will” and “determinism” – must meet to serve in this capacity.

“Free will” (here sharply distinct from “free action”) is thus broadly taken to refer to some description of the use of the individual human mind that collectively fulfills the following necessary conditions: (i) some minimum criteria of rational competence of a human mind about the known or believed contents of mind (propositional or phenomenological) relevant to this use, (ii) some plausible situation of choice or decision between known or believed logically exclusive propositional alternatives relevant to (i) (usually ultimately actions but perhaps only mental items), (iii) some presence of control or ownership of (i) and (ii) that ties those factors into a metaphysically coherent concept of autonomous agency by a human mind of its contents, (iv) a lack of significant external non-mental coercions about the overall situation of (i)–(iii) (thus mooted most direct questions of free *action*⁴) as well as a lack of significant external coercions that might result in a psychological manipulation of an agent, and (v) sometimes involving moral overtones of the use of free will via (i)–(iv) to involve matters of responsibility for such use.

“Determinism” cast as free will’s incompatibilist villain is thus broadly taken to refer to some logically possible set of circumstances that conflicts or significantly interferes with any or all of (i)–(v). Many of the types of determinism referred to above are often taken to be sufficient to do that in one way or another. This (taken as) fact begs to question why one – causal determinism – ought to be taken as the most fundamental one.

A lengthy monograph would be needed to work out this question’s answer in any sort of detail. Here, brief caricature must do. Note that many types of determinism undermine (ii) and (iii) by interference with individual human agency – God, social groups, truth-fixity, and causality – and thus undercut or by-pass the role of the human agent as illusory or inconsequential. There is some empirical evidence that many people express alarm about the diminution of agency in such circumstances.⁵

Aside from causal determinism – to be dealt with in a moment – the other listed forms of determinism have a form of common disadvantage as serious contenders for being the target of the compatibilism–incompatibilism debate. This is that they are *external to*, or perhaps, inappropriately *externally formative of*, individual human agency as expressed by (i)–(v) above. A Calvinistic God might extinguish individual human agency, oppressive socialization might overwhelm it, and considerations of logic might confine it, but always in some *extrinsic* sense to any question as to how minds function. To make the case for any of these involves some argument that takes place *outside* of what is usually taken in ordinary circumstances to constitute individual human agency. Thus we should argue for the existence of an entirely

providential God, produce cogent data for social dominance over individual autonomy, or show how logical alternatives imprison us as to impinge on our autonomy, *and then* show how these entail revision of how we usually think about individual agency and human action.

Causal determinism, on the other hand, involves some account of *the very ground* of how individual human beings seemingly engage – but might ultimately *fail* to engage – in the activities (i)–(v) without essential reference to any external entities or issues, even if it also might serve as an instrument to explain how such external things might work to undermine free will.⁶ Thus causal determinism maintains, as expressed through incompatibilism, that its truth somehow rules out at least some of those factors of agency as an *intrinsic* account of how we behave. That means we must now investigate how and why causal determinism would do that.

1.2 What Is Causality?

This is a hard question to answer mostly because there is not even general agreement as to how the nature of the answer should look. What we can do however is to sketch the most serious answers in terms of what they attempt to do with respect to capturing what we typically mean by saying something is a matter of cause and effect. What all accounts of causality have in common is that there is in human experience some strict connection between many asymmetrically temporally connected earlier-to-later phenomena. Most accounts emphasize the rigidity of this connection as a form of one-to-one relationship (whether event-token or predicate-type) while others loosen it to constrained one-to-many relationships, usually by references to probabilities of subsequent event- or type-effects relative to a given cause. Because the form of determinism that stands behind the compatibilism/incompatibilism distinction typically relies on the first kind of one-to-one rigid relationship, I will ignore probabilistic accounts.⁷ So, if human will and actions are intrinsically subject to such strict one-to-one connections, then doubts arise about whether the factors (i)–(v) above may be fulfilled, particularly the control/ownership criterion of (iii), since the dominance of causality seems to swallow any question of an individual's autonomy apart from that.⁸

Some accounts of causality are overtly *metaphysical*, trying to capture the real nature of causality based on some confidence of available knowledge about it, while others are versions of a more distinctly *epistemological* approach with respect to that available knowledge, and vary in holding just how cautious we must be in our resultant metaphysical commitments.⁹ I will simply refer to the former accounts of causality as *M-causal accounts*, and the latter as *E-causal accounts*.¹⁰ Because these approaches offer strongly different though at times logically compatible postures with respect to the phenomena typically described as exemplars of “cause and effect,” these two accounts are not, like incompatibilism and compatibilism, exclusive of possible positions. My own final view will for example combine them, though lean mostly toward an E-causal account. There are two prominent forms of these two accounts that currently dominate discussions of incompatibilism and compatibilism with respect to the issue of determinism. Next I will outline each and criticize them as unsatisfactory accounts of determinism, and as well I will then offer my own as a compromise view.

1.3 An M-Causal Account of Determinism

The most well-known M-causal account is due to Carl Ginet (Ginet 1990) and Peter van Inwagen (van Inwagen 1983), and is based on the assumptions that (a) causality underlies the laws of nature and (b) human actions are entirely explainable by the laws of nature.¹¹ While neither Ginet nor van Inwagen give precise accounts of the nature of causality, clearly

they agree that causal determinism (if true) undermines human agency by constituting the basis for all human action.

The clearest use of this kind of account is van Inwagen's Consequence Argument, which has appeared in several forms. For the purpose here, we need only a simplified version:

If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us. (van Inwagen 1983, p. 16)

This language is sufficient to see that causal determinism of this sort conflicts with free will's conditions (iii) and (iv) above ("it is not up to us"), and thus nullifies any importance of (v). The basic idea is straightforward: if laws produced the way the world was before our birth and also produced for each of us our own individual human nature as part of that, then as part of nature we always and only have at any time one single thing we can think or do – what is causally necessitated by the laws. Therefore if causal determinism is true, and conditions (ii) and (iii) are necessary for free will, then free will cannot exist. This is how most versions of incompatibilism are understood in contemporary free will literature.¹²

This interpretation of how causal determinism would function in the world and human nature to eliminate free will has been harshly criticized from many different perspectives.¹³ We need focus on only two general ones.

The first criticism is that, while the argument may capture an accurate depiction of the way human nature works in terms of causal determinism, the free will conditions (iii) and (iv) as stated above are erroneously construed by the argument to conflict with the truth of determinism. So human beings may at once be deterministic mentally and physically and yet also possess free will in the sense that at least our present actions *are* "up to us." In effect this is to advocate some compatibilist view of free will over incompatibilism.

The second criticism holds that causal determinism as interpreted by the argument is just false. Quantum physics has shown that there are laws of nature that do not work by any plausible concept of cause and effect at least at the subatomic level, and thus indeterminism is true both universally and as constitutive of human nature.¹⁴ Therefore not all laws of nature are deterministic.

While these are both relevant criticisms, I will focus only on the second, since the first denies that determinism even of the strongest M-causal type conflicts with (ii) and (iii). My own view, as I will argue, maintains that it is plausible that human nature is causally deterministic even if not all laws of nature are deterministic. My version of determinism may not salvage the consequent argument as it stands, but it may save enough of its motivation to undergird incompatibilism, as the consequent argument intended.

1.4 An E-Causal Account of Determinism

Unlike M-causal accounts, E-causal accounts exhibit as a group greater caution about the nature of cause and effect based upon our limited knowledge of it as a kind of one-to-one observed phenomenon. A prominent view in more recent free will literature is one advanced by Helen Beebe and Al Mele as informed by the work of David Lewis, which in turn is based on Humean skepticism about our knowledge of the nature of causal relations.¹⁵

Beebe and Mele's E-causal view also depends upon the idea that determinism is true only if the laws of nature are entirely matters of cause and effect. However, what constitutes a

deterministic description of the world and its laws is inherently time-dependent, based on (a) our observations of the different kinds of event regularities up to any point of reference and (b) a holistic summary of these various kinds of events construed as the laws of nature. The point of (a) is to concede the Humean skeptical point that cause and effect relationships are no more than our observations of the constant conjunction of kinds of events, and so may not be projectable into the future. The point of (b) is that what constitutes the collective of the laws of nature based on (a) has to be justified as properly explanatory as a bundle of laws at any given time. Lewis's account of (b) – called the Best Systems Analysis, and endorsed by Beebe and Mele – is that the laws of nature form the best system that satisfies norms of explanatory comprehensiveness and simplicity.¹⁶

Epistemically Beebe and Mele's account of determinism is a minimalist one, and in fact many M-causal accounts of determinism are compatible with it as explanatory of how we *know* the laws of nature at any given time. However, M-causal accounts typically hold that, if our knowledge of the laws are correct at any time based on the past, then we may be confident that those laws will hold into the future. Beebe and Mele deny this: it is both logically and metaphysically possible that laws that have held up to a certain time may deviate in the future from the forms of constant conjunction that justified them up to that point.

This results in a form of determinism that appears to evade van Inwagen's consequent argument by holding that the laws of nature and the past cannot fix the future beyond any given point of time that they have held. Therefore, even if human nature is deterministic in this sense, it is both logically and metaphysically possible that a human agent *might* fulfill conditions (i)–(v) above at any given time, thus resulting in a form of "Humean compatibilism."¹⁷ This E-causal type of determinism is singular in the respect that it undermines the motivation behind the compatibilism/incompatibilism distinction if it can combine determinism and conditions (i)–(v) coherently – there would be no reason to think that a concept of indeterminism would be of any relevance to thinking about free will at all, for E-causal determinism delivers all that any indeterminism could, but without sacrifice of believing in the rule of physical laws (at least up to any given point of time).

While Humean compatibilism is an ingenious end-around of the compatibilism/incompatibilism debate, it has inherent tension with point (iii) about providing a coherent metaphysical account of how agents could *control* choices between alternatives. On the one hand it countenances Humean variability of choice at any given point of time due to skeptical uncertainty about future courses of events; on the other hand it has by embracing an epistemically based minimalism no resources to assure the crucial element of control of such choices.¹⁸

So now we have two prime examples of very differently justified views of determinism. One is based on the fact that cause and effect relationships are so justifiably grounded in experience that we can posit some sort of metaphysically necessary account of those relationships across all time by referring to universal laws – M-causal determinism. The other more cautiously is descriptive of such cause and effect relationships up to a time bundled together as an explanatory system of laws that might deviate from that time-fixed description in the future – E-causal determinism.

There is one assumption that binds both of these very distinct views together: determinism is the belief, epistemically time-fixed or not, that human nature would be a function of the *entirety* of the laws of nature, and any idea of free will via (i)–(v) must play off that fact. I think that's false. Now I must make my case for a different way to think about determinism, but a way that preserves it as the basis for the compatibilism/incompatibilism debate.

2 A Relativist Concept of Determinism

So far we've seen that what threatens free will's condition (iii) in M-causal determinism – causal necessity accounts for all human action, leaving no room for any other control or ownership by a human agent – might be avoided by E-causal determinism. So M-causal determinism upholds incompatibilism, while E-causal determinism does (or might) not. However, there is another option – a form of causal determinism that shares some strengths, and alleviates some weaknesses, of both these types. This is a kind of determinism relativized to systems of events – R-causal determinism.

There is a dual motivation for this position, partly drawn from M-causal accounts and partly drawn from E-causal ones. The first is that one weakness of M-causal accounts concerns the seemingly well-secured fact that quantum indeterminacy is a well-established law of the land: at least at sub-atomic event realms, metaphysical indeterminacy may well be true.¹⁹ But second, the strength of E-causal accounts is that there are empirically relevant non-quantum systems of events that appear, even assessed by rigorous standards of measurement, to qualify as deterministic. Something is going on here about how we should rationally assess the goings-on of the world as somehow plausibly including both deterministic and indeterministic accounts. A relativism of explanatory context would do the job, based on what kinds of systems of events were involved. To be blunt, the collective motivation of these two observations is pragmatic – we seem to need to revert to indeterminism or determinism based on how we best wish to explain this or that phenomenon. Want to explain the Compton effect? Indeterminism reigns. Want to explain how billiard balls knockabout on tables? Determinism is the best bet. Context of which kind of explanation is pertinent is everything. Nancy Cartwright's and Bas C. van Fraassen's work is a key inspiration for me here.²⁰

Which brings us to the key question of free will – which context of these two forms of explanation best fits how minds operate, especially in terms of criteria (i)–(v)? Of course indeterminists – libertarians about free will mainly – will insist that the global context of explanation holds the day – quantum theory demands it. But determinists should counter that even while the indeterminism of quantum events might be true at that level of reality, the brain – and mind – might well function by means of bio-electro-chemical processes that obey deterministic laws. The jury is out on which context is the final right one in the arena of how best to explain human action. However, if we hold the jury to standards of preponderance rather than proof beyond reasonable doubt – there is little to question that the majority of human mental and resultant physical actions are the result of something very much like cause and effect – the countless daily mindless habits of normal life are testimony enough. It is from that perspective that I will operate. It is more likely that determinism of human action is true than not, and thus I will attempt to define a determinism of human nature largely motivated from that likely perspective.²¹

To begin, we require some plausible ontology of human nature that offers the possibility that it might be described deterministically, but also might not should human nature be indeterministic. The best candidate is an event ontology, since it seems possible to describe both the physical and mental sides of human nature as two types of events.²² We begin with a brief definition of a physical event, put forward in purely empirical terms:

***Physical event:** any occurrence that can be measured (to an arbitrary degree of precision – ADP) in four dimensions of 3D space and 1D time.*

Several observations are needed here.

Events are carved out of the universe by us – as objects of scientific inquiry they do not individuate themselves. It is in human interest about portions of the universe where the existence of physical events as individuated entities begins.

Therefore, *where* and *when* events occur as measured are always functions of human interest. So one aspect of the measurement qualification ADP above concerns primarily subjective matters – what do we think constitutes sufficient carving to satisfy those interests, and how accurately must we measure to achieve that sufficiency?

The universe also makes objective contributions here, however. Part of that is constituted by the nature of materials that make up measuring devices themselves – we can make as accurate measuring devices as needed only if we have adequate materials to embody our needs, both in terms of spatial and chronological accuracy. While human ingenuity has produced impressive instruments of measurement of space and time, there also seem to be inherent limits of accuracy imposed by nature itself, e.g., the Planck length.

The other part of nature's objective contribution is – well – the nature of nature as revealed to us. While sometimes nature reveals very unusual events that strike our interest, most often nature reveals repetitive and uniform types of events across space and time. For the purposes of defining determinism, it is to these latter types of events we will turn.

In summary, physical events are objectively measurable – by anyone – and everyone so interested – as found objectively in nature. Physical events then are *objective* and *public* as measurable.

Examples will help in advancing the argument. I will use a familiar one – a certain kind of a billiard ball collision.

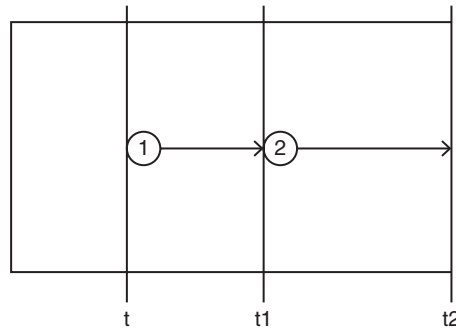


Figure 3.1 A Simple Example of Causation.

Here by stipulation ball 1 is struck such that at t it moves on a path parallel to the sides of the table so that it strikes ball 2 in the middle at t_1 , and so that ball 2 then moves along the same path as ball 1 to the end of the table at t_2 (all described motions and points of contact measured to an ADP).

Experience shows that, even if balls 1 and 2 are transtemporally deemed as the same balls used, this kind of occurrence can occur at different values of t labeled sequentially. Substitute balls 1 and 2 for similar balls X and Y , similar events of this kind can occur at two locations (two tables) at the same time. So Figure 3.1 describes a *kind* or *type* of collision – if the intervals described by t to t_1 and t_1 to t_2 obtain anywhere, then this is the same *kind* or *type* of collision.

Now let's assume we are describing any such collision. There are background conditions that must obtain for such a collision to occur, and such conditions are not easily delineated exhaustively. But many can be specified, and all can be placed into two categories of necessary

conditions. For example, one such set of conditions includes (ADP) facts such as: the table must be level and surface smooth, the balls similarly constituted and constructed, the first ball struck so as to move consistently with the description of its parallel motion, no one is purposely vibrating the table, etc. But a second set of conditions must also obtain: gravity must hold the balls to the table and not fluctuate, no earthquakes must occur during the t - t_2 interval, etc. The first set of conditions are ones capable of human control; the second are uncontrollable – *both must be the same (by ADP) for the kind or type of collision to occur*. I will call these *standard conditions* or SC for some such kind or type of occurrence as in Figure 3.1.

So any such occurrence as represented by Figure 3.1 along with SC is here taken as iconically representative of this kind or type of event.

Now introduce specific measurements of a particular collision. Say that (under something like tournament or even better laboratory conditions) such a particular collision occurs at Faye's Pool and Gruel in Green Bay, Wisconsin at table 2 (even if similar things happen at table 1 at the same time) at $t = 11:00$ a.m. sharp CST February 10, 2022, and progressing for interval t_1 ending = 11:00:01 and interval t_2 ending = 11:00:02. SC hold throughout. Now we have space and time measurements specific enough to label these occurrences as respectively specific events E and F (Figure 3.2):

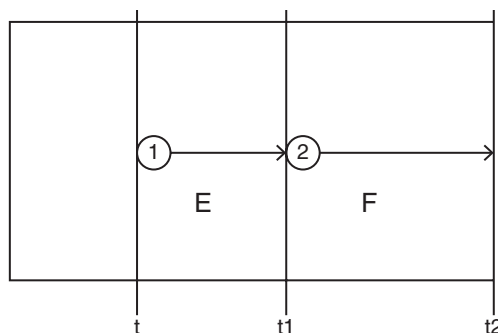


Figure 3.2 Example of a naming convention for causally relevant contiguous events.

Notice that we could have defined the entire collision from beginning to end as a single event. Here it is our interest in defining cause and effect that justifies subdividing this occurrence into two events standing in such a relation.

Now, whereas space measurements may replicate – an inch here is the same inch there – rigorous time measurements are ideally unique – one time interval occurs and never reoccurs. So once E happens and is labeled, E may never happen again. Thus event naming is by an ideal form of *rigid designation* – once a name is used for some one event, it may never be reused. If, for example, balls 1 and 2 are repositioned even at the same day and place as E and F's occurrence but beginning at 11:00:30 move the same as E and F under the same SC, they are not the same events – they must be given another name as only uniquely described by their times – maybe G and H. Physical event names are ideally unique, and thus completely individuate physical event occurrences.

Now for some consideration of possibilities, which is necessary for defining causality, and thus determinism. Consider this scenario based on the presumed occurrence of E (Figure 3.3):

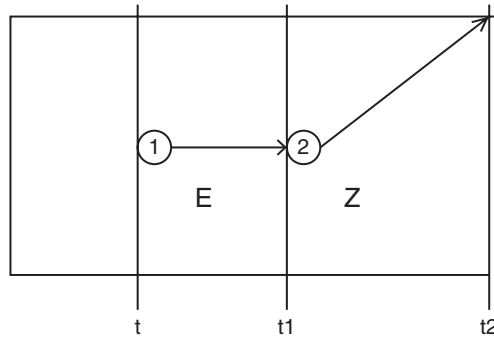


Figure 3.3 Illustration of a logically possible set of contiguous events.

Here E occurs, thus at the same time and SC as E. Now, let's say that Z then occurs. What does this mean?

First, since ball 2 is involved, Z has some overlap with what occurred with F – they both involve the same ball 2, and at the same time and place as involves their collision.

Second, this raises a technical issue – why should we think of E and Z to be connected as E and F were presumably connected? Namely, how do we know that E and F were actually connected events, and why should we think E and Z might be?

The connection of events is indeed a technical and conceptual issue. It is logically possible that there is some perhaps undetectable and immeasurable interval between E and F just as there may be between E and our hypothetical Z. How may we assure that doesn't occur in either case?

The answer is a stipulation about the ADP circumstances of measurement: we stipulate that the ideal singular moment of the end of one event constitutes the same singular moment of the beginning of the next event. That way there are no intervening moments of time to separate events, even ideally.²³ This utilizes specification of particular points of the (infinitely dense) *continuity* of time and space to assure the *contiguity* of events in time and space, and thus to undergird claims about how reality works through space and time interpreted as events. Descriptively we may say for each scenario, “E occurs, ball 2 is struck in the middle, and then X...” describes the subsequent event X then happening, taking “ball 2 is struck in the middle” as describing the moment and place that E ends and the subsequent event X (as a variable) begins.

So Figure 3.3: E occurs and then Z occurs, where ball 2 is truck in the middle and then Z proceeds on a straight line toward the corner pocket. E and Z are thus continuous in space and time and contiguous in space and time as were E and F. This poses the question: given E – and SC hold throughout the interval – is Z possible?

Ordinary experience says no, Z is not possible given E. But – can we *imagine* that Z occurs given E – and SC hold? The answer is clearly yes!

Such a case is one of what is *logically possible* (LP). We can consistently imagine that E occurs, then Z does, despite the fact that we hold SC constant. All of fiction – some less or more imaginative as ranging from historical fiction (less) to fantasy (more) – involves the same kind of logical possibility. Anything that attempts to describe logical contradiction – E occurs (and thus ball 2 is struck in the middle) and then ball 2 both moves in some way, but at the same time does not move in any way – is an *impossibility*, and cannot constitute a consistent description of a subsequent event. Alice in Wonderland describes very unlikely sequences of events, but they are entirely LP.

So Z is in fact LP given E. So is T – where T is that ball 2 is struck in the middle, and then it immediately turns into a toad (T) and then jumps off the table!

Now, more seriously, let's imagine another event given E–D.

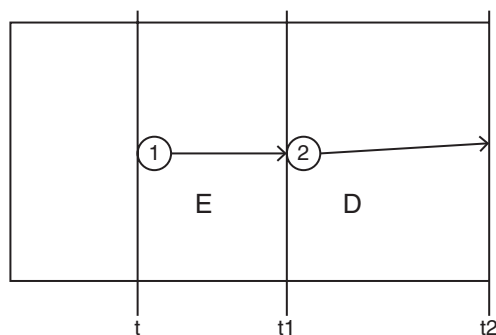


Figure 3.4 Illustration of logically possible contiguous events that are metaphysically (real world) plausible.

What is D descriptively in Figure 3.4? To secure contiguity, D is that ball 2 is struck in the middle, but then moves across the table on a path that is $1/2$ of a degree deviant from F – a measurable difference that demands a different event label. All along SC have been assumed, as under something like laboratory conditions where an experiment has been conducted.

D is clearly logically possible given E, assuming SC. But what if D (for Deviance) is actually observed in a lab where E is established by measurement (including the collision of course as required for E)?

In real labs with very precise instruments, the measurement of D given E under SC would be a puzzle. Why didn't the expected F occur? There would be inquiries – F was expected for good physical reasons that applied to the experimental conditions – and thus D should not have happened. Perhaps it will be found that just after t_1 there was an earth tremor, or that a lab assistant had unknowingly dropped a pretzel salt crystal from his beard on the table that lodged right under ball 2. These would constitute violations of the assumption that SC held throughout, and would thus also constitute the grounds for explaining why D followed E.

Either case – of E and F given SC or E and D given SC plus NaCl in the salt instance – shows that these events are what we expect in the real world *given by hypothesis* the laws of physics that govern the movements of billiard balls. This provides us another idea of possibility relative to what we expect to happen in the real world *given* our tentative understanding of physical laws and expectations about relevant SC-like circumstances. We will call this *metaphysical possibility* (MP). So, F is metaphysically possible given E under SC, and D is metaphysically possible given E and SC plus NaCl.

However, the fact that these MP results are aligned with particular event descriptions under particular SC-like circumstances strongly suggests that these are more than just possibilities, but real physical *necessities*. With such an idea in hand, we may now precisely define an instance of *cause and effect*:

E causes F: Given that tentative strictly governing laws hold and that E occurs and SC also hold, then of all LP events X that may be thought to occur after E, F and only F is the LP event that is also MP.²⁴

Note that we could show that this definition generalizes. We could define *E causes D* only substituting “SC plus NaCl” for SC. Causality occurs within a system of events under specific circumstances assumed to be governed by strict laws and is a relation between two unique events.

It's important to see here that these descriptions and definitions follow the lead of well-grounded empirical science as it is practiced in laboratories. When physicists conduct careful

experiments using items like balls 1 and 2, they control conditions as carefully as possible to fulfill their posit of SC-type conditions on the control side, but cross their fingers that non-controlled SC-type conditions obtain throughout the experiment. And frequently they do. But even when they do not, it does not appear that causality is violated in such experiments – only scientists' expectations about the preservation of posited SC.

As observed before, the universe may be carved up into events as we see fit. We have found that some such carvings at the subatomic level yields laws and SC-sets that might not in fact produce strict cause and effect relationships between events, namely quantum events, and we will describe such relationships later, logically based off what we find with describing causally related events. However, some carvings of some events such as billiard balls yield very stable descriptions of event relationships as causal, and this suggests that at least some systems of event carvings are describable as causal – even *entirely* causal. If we limit our attention to the movements of balls on the table during an entire billiards game, it appears all such are causal (for example). So, even if some systems of events are not entirely causal, as quantum systems seem not to be, some others are. And not just systems like billiard games, but much more complex ones like the weather, biological systems like plants and simple animals, but maybe even more complex varieties of biology: the human brain and mind. The causality of systems may be a function of what is described, and it may be that we humans at the most fundamental level might be subject to causality.

So – ascriptions of causality – even complete causality – might be *system-descriptively relative*. This is the basis for **R-causality**, founded pragmatically in the actual rigorous practices of some sciences on the basis of tentatively entertained laws governing systems of events studied by those sciences.

Let's now represent an R-causal relationship as it evolves through time on a spacetime diagram.

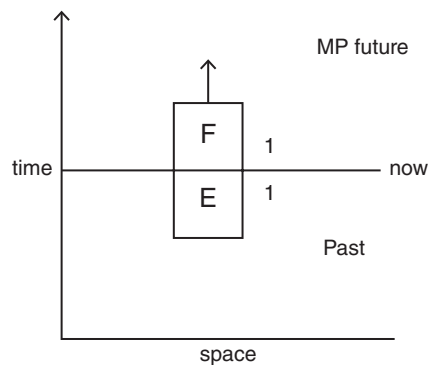


Figure 3.5 A simplified spacetime diagram of Figure 3.2 with a present moment as the point of contiguity between events illustrating causality through time.

Here the system constituted by E and F are presented as evolving through time on the vertical axis and as they occupy space on the horizontal axis (3D measurements collapsed to 1 and further simplified). The moment labeled “now” obviously is t_2 from Figure 3.2 – the end of E that also serves as the beginning of the next event. F is given as the next event, indicating that this diagram represents in the background that SC strictly hold throughout. E has thus occurred in the local past, and F will occur in the local MP future, and in a strict one-to-one relationship of particular event to particular event.

As a picture of causality functioning through time from past into future, Figure 3.5 shows very clearly a feature that is touted by most versions of M-causality: given an actual past and causality, the future is fixed. This diagram explicitly shows how this is the case in a particular instance of causality with reference to a carefully delineated local past and local (MP) future. It thus literally represents the definition **E causes F** presented above.

Now we are positioned to give a definition of *R-Determinism* based on this concept of R-causality:

R-Determinism: A system of events wherein any and all such events exhibit only R-causal relations between any two spatiotemporally contiguous events.

So what if some system of events does not qualify as R-Deterministic? Then regarding Figure 3.5 *entirely* as a LP way to depict one way events may evolve into the future – LP₁ – then we may draw a contradictory diagram LP₂:

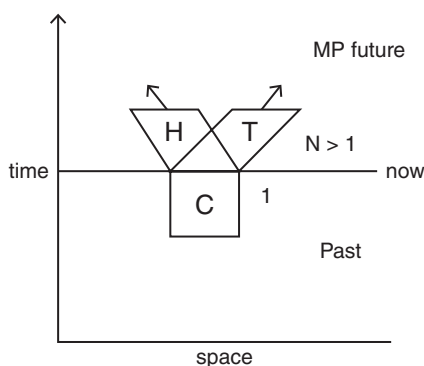


Figure 3.6 A minimal logically possible spacetime illustration of non-causality in contrast to Figure 3.5.

Here we have an event C that ends at “now,” but unlike an R-causal case, the MP future allows at least two distinct subsequent events to occur, H and T. For convenience, these labels represent another familiar physical event case – a coin flip. C then is a coin flipped and ends when “now” occurs, which is the ideal moment that the flip turns H heads or T tails. The label $N > 1$ is a reminder that such a process might allow more than two outcomes (maybe landing on edge in this case).

Figure 3.6 thus is representative of a relationship we may call constitutive of *R-Indeterminism*:

R-Indeterminism: Any system of events wherein it is not the case any and all such events exhibit only R-causal relations between any two spatiotemporally contiguous events.

Note that this definition is consistent with calling a system R-Indeterministic even when it includes some events that sustain R-causal relations; no matter how complex a system may be in terms of events, the presence of just one R-Indeterministic relation between MP events relative to some past event makes that system R-Indeterministic.

It is important to stress Figure 3.6 represents a *real* cross-temporal one-to-many relationship of MP and *not* just some *epistemic* probability. Even if coin tosses should turn out to be R-causal – and this very well might be the case – we still use statistics and probability to project possible outcomes of coin tosses in epistemic fashion.

However, now we have a way, by something like Figure 3.6, to represent how quantum events evolve into the future as plausible examples of R-Indeterminism. Between R-Determinism and R-Indeterminism we thus have an exhaustive way to talk about how any system of events works through time in the real world, and allowing that different systems might be legitimately describable by one of these, depending on what is demanded by rational use of science – or philosophy.

We now just lack one more step to complete this toolkit – expanding everything here to include *mental events*. So first, a definition:

Mental event: any occurrence that can be only be subjectively measured by a self-conscious mind (to an arbitrary degree of precision – ADP) in the one dimension of time.

This is a rather unusual definition, admittedly. However, it has clear advantages over (say) a more out-rightly philosophical one. (i) It is (almost) completely empirically based, relying on the same concept of empirical measurement as does the definition of physical events. (ii) It concedes the point of a more philosophical definition by allowing the essential subjectivity of this kind of measurement – only a given conscious mind, aware of its contents as well as the correlation of those contents to measurement by means of an ordinary chronometer, is able to accomplish this kind of measurement. Only *we* in our own instances of self-consciousness can say, to an ADP, that a particular mental content occurs during durations of clock-time. Therefore, mental events are *subjective and private*, as opposed to physical events as *objective and public*.²⁵ (iii) The uniqueness of rigid designation is thus as secured for mental events as it is for physical events – the times of mental events are the factor of individuation in all cases (along with references to a given mind, perhaps, as an added indexical reference should different minds occupy different bodies in the same temporal reference frame with comparable mental content).

So now we may produce representations of R-Determinism and R-Indeterminism for *minds* as mental events mirroring all aspects of those ideas based on physical events, with just one technical adjustment in addition to an appropriate change of example (Figures 3.7 and 3.8):

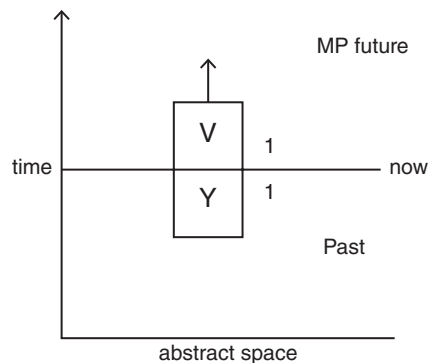


Figure 3.7 An abstract illustration of a causal relationship between mental events through time.

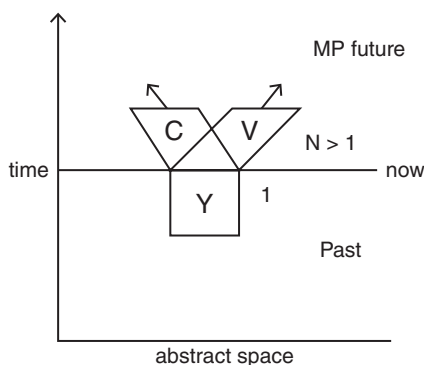


Figure 3.8 A minimally logically possible abstract illustration of a non-causal relationship based on Figure 3.7.

The labels are descriptively for a given mind in two LP scenarios of Y related to two distinct MP futures in respective diagrams: Y represents a mental choice of LP frozen yogurts offerings (while visiting a shop) during a time interval encompassing Y, involving LP C chocolate and V vanilla, where those mentally presented LP choices during the deliberations of Y are at least “close” in terms of desirability and downsides and thus are real choices as consciously entertained during the span of Y (for mental events this same descriptive basis for the rigid designation of Y across both diagrams stands in as well for *one mind* for the assumed stable SC conditions of physical event descriptions, as required for the possibility that mental events do not have a SC physical foundation according to metaphysical dualism²⁶). Figure 3.7 represents the fact that one LP – the V choice event – is MP caused to occur (if it should work out that way in the real world), and Figure 3.8 represents the fact that both LPs of C and V are MP possible (at least – maybe a twist of them is also possible – thus the $N > 1$ label). So Figure 3.7 represents a mind that works at that time by (a minimal instance of) R-Determinism, and Figure 3.8 represents a mind that works at that time by (a minimal instance of) R-Indeterminism.

The technical adjustment required is that the axis of space in the ordinary spacetime diagram for physical events is replaced by a form of *abstract* space merely to represent an extra dimension so that mental events might be viewed in a 2D fashion parallel to ordinary spacetime diagrams. The justification for this is similar to how imaginary and complex numbers are represented in imaginary space – this is a way to represent some real quantity (as is the time of mental events on the Figures above) in a coordinate space so that it might be visualized.²⁷

We now have all we need to construct basic concepts of incompatibilism and compatibilism as based on R-Determinism and R-Indeterminism.

There is one large advantage of R-determinism over M- and even more cautious E-forms. Whereas both of those crucially rely on the central concept of *law* in one way or another in order to define determinism, the R-construct is based only on tentative *lawlike* epistemic grounds (of experienced types of event regularities even if asserted to be “laws of nature”) in order to justify the expectations behind the event-based descriptions of the relationship of causality and its logical opposite in given cases. The asserted determinism of a system, or of a certain type of system, is thus always a function of what system is described, and is always epistemically tentative in terms of its evidential basis and what might be observed in the future (thus reflecting some of the skeptical wisdom behind E-causality). We have good grounds at present for saying that a deterministic view of billiard games is justified; we have

good grounds at present for saying that an indeterministic view of quantum events is (well-enough) justified. But what is up for grabs at the moment is whether the crucial system of events that constitutes the basis for human choices is one or the other of these kinds. Still, conceptually R-determinism and its opposite provide an empirically adequate, clear, and logically coherent foundation for further ruminations on free will based upon the traditional incompatibilism/compatibilism distinction.

Earlier it was argued that belief in causal determinism intrinsically undermines free will in ways that other plausible forms of determinism do not. The R-Determinism of Figures 3.7 and 3.8 help to show why this might be case. In Figure 3.7 a mind is causally fixed to a “closed” local MP future that is unavoidable given its cause, which by hypothesis is a mental event description of a decision process between two LP alternatives entertained by that process conceived as such. In Figure 3.8 that same description of mind is not causally fixed, resulting in an “open” local future constituted by at least two MP possibilities that represent the mind’s content about the LPs.

Two crucial necessary conditions of free will cited earlier are relevant here: (ii) some plausible situation of choice or decision between known or believed logically exclusive alternatives of content relevant to [presumed mental competence], (iii) some presence of control or ownership of [that competence] that ties those factors into a metaphysically coherent concept of autonomous agency. The truth of Figure 3.7 would render any epistemically possible beliefs about those factors (ii) that an agent has about the logically exclusive alternatives (C and Y) effectively moot – only one would obtain from that particular event, and causality would exclude the other as not MP. This would also show that the causality that led to the locally future event would undermine (iii) – there can be no control of what happens outside of causal control, and thus no conceptual room for human agency besides. It would appear that R-Determinism of a mind then stands as a potentially sufficient reason to claim it as *incompatible* with free will. (This captures much of the basis for M-Determinism’s incompatibility with free will as expressed by the Consequence Argument but without having to embrace universal causal determinism, thus avoiding a confrontation with the truth of quantum indeterminism.)

Figure 3.8, however, would stand as a basis to deny both of these claims, thus making it minimally sufficient as a basis for claiming that it represents fulfilling conditions (ii) and (iii) for free will. Thus it would stand as a LP positive representation of a real state of incompatibilist free will, as far as it goes in opposing the implied negative incompatibilism of Figure 3.7.

But these diagrams also help us understand at least some compatibilist motivations, especially since they encompass (as contradictory but exhaustive of) all LP views of how minds might evolve into a MP future. If these two depictions of MP diachronic minds are somehow conjointly *not* definitive as to how minds might be unfree or free (7 or 8 by incompatibilism), but by some adequate account of compatibilism are nonetheless plausibly free in some instances, then something about the *descriptions* of a mind – such as the mental content of Y in both diagrams – are more determinative of how (ii) and (iii) might be fulfilled than matters about future MPs. Contemporary accounts of “reasons-responsiveness” for example attempt to show how a certain descriptive mind-set might by Figure 3.7 R-causally result in a decision yet show LP close-world counterfactual sensitivity to reasons that nevertheless does not at all assert that the MP future of Figure 3.8 is relevant.

As well, Beebe and Mele’s E-Determinism, for example, allows that minds might up to any given time descriptively behave as Figure 3.7 and thus be assessed as deterministic, but at any given future time – and any given past time as counterfactual – might (have) behave(d) as Figure 3.8. This configures a form of compatibilism (with a deterministic mind up to some point of that assessment) that countenances at least future LP Figure 3.8 instances of MP

indeterminism or that same possibility as counterfactually LP posited against any time that R-causality actually occurred. E-Determinism thus is purely epistemic as far as causal relations are concerned: they are only justified by constant conjunctions of event types up to any given time yet *may* fail at any future times, and counterfactually *might have* failed at any time. The potentially metaphysically *accidental* nature of E-causality is explanatorily central to such compatibilism.

3 R-Determinism and the Free Will Problem: A Way Forward

Let's review. Causal determinism has long stood as the axis around which the free will problem turned, and for some good reason: if causality stands as the final explanation for how minds work, then that fact arguably crowds out the venue for agency as the only explanation for how minds controllably work. M-Determinism, by assuming the universal sway of causality through the dictates of universal laws, does exactly that, especially along with the force of the Consequence Argument. However, the rise of quantum theory in its indeterministic guises stands as an empirically stout challenge to that position. E-Determinism, by similarly assuming the possible absolute nature of determinism as expressed by known laws, but allowing that those laws might change, or might have changed, at any time, makes room for a consideration of alternative possibilities consistent with indeterministic considerations, and so possibly consistent with conditions of free will. However, the price paid is an implausible concession to skepticism about how the world works – that it is possible that the world or parts of it does not work by a regularity of final and absolute deterministic law, but only by some sort of magnificent array of accident.

By defining a form of causality empirically fixed to contexts of event-based explanation as practiced by science and based on tentatively posited laws of such systems, however, we can avoid some of the difficulties of M- and E-Determinism, yet salvage some of their strengths. This is the comprehensive and pragmatic justification for R-Determinism.

So if R-Determinism of minds is generally true – and that is in part an empirical question that might be answered by neuroscience in future – then there is a legitimate basis to claim that incompatibilism might also be generally true, plausibly extended into the future, and even apart from the Consequence Argument. As well, should R-Determinism be true of human nature, that would serve as an adequate companion for any durable form of compatibilism independent of that truth.^{28,29}

Notes

- 1 A typical logical artifact of interdefinition by negation is that one of the binary definiens is ambiguously logically open-ended, a converse as “non-red” constitutes an open-ended set of possible colors or combination of colors against “red.” My point here is that most students and even distinguished scholars of free will take compatibilism as basic, and incompatibilism then interdefined by negation off that. While that yields incompatibilism as not-compatibilism, that would be like in the color analogy taking “non-red colors” as basic and its logical opposite “[a] non-non-red color” as interdefined off that. I think the main reason for this state of affairs is that compatibilism has become the most popular general view (Bourget and Chalmers 2014), and thus the minority view is ad populum dubbed “incompatibilism” off compatibilism, but that also unfortunately disguises the logical fact that compatibilism, and not so much incompatibilism, is an open-ended, many splendored general view, constantly susceptible to possible revision and expansion as a set of accounts of free will. I should take note that all these remarks are classically logical simplifications from the point of view of Kristin Mickelson in this volume

- (Ch. 4), who sees the compatibilism/incompatibilism debate as much more nuanced, though I also think that my distinctions could be embraced in her careful analysis for the purposes here.
- 2 Some authors (Mawson 2011) take the question of incompatibility to be with moral responsibility rather than free will. Since the motivation for this take on the debate is plausibly supervenient on that same motivation for positing free will as incompatible with determinism (Fischer's semi-compatibilism takes exception to this (Fischer and Ravizza 1998)), I'll stick with the focus on free will.
 - 3 Or as argued in the first footnote so as not to derail the main text, to define "incompatibilism" as the main term!
 - 4 For clarity this follows the distinction outlined by Gary Watson (1986).
 - 5 See, e.g., Vohs and Schooler (2008).
 - 6 Eighteenth-century deism, for instance, could easily invoke causal determinism as the mechanism for how a providential God might foreordain everything. That was its strength: Newton's and Calvin's views were compatible after all as educated people grappled with the problem of faith and reason. Nevertheless, deism is an externalist factor as I have argued here because God as predestinator is required to fix fates even if causality is invoked as its mechanism.
 - 7 Probabilistic views of causality in fact are usually motivated by belief in either metaphysical or epistemic indeterminism, and involve complicating matters of luck that may well undermine the force of incompatibilism independently of the role that causal determinism might play intrinsically in its role in that view as argued here (see work on preemption and "fizzlers" in Hitchcock 2021). I should also point out that even some libertarians rely on the one-to-one sense of causality; i.e., Kane (2004) and O'Connor (2009) as event- and agent-causalists (though recently Kane has denied he is an event-causalist, Clarke, Capes, and Swenson 2021). Both these views use causality to produce actions in the more usual sense of causality as a one-to-one relationship, even if they also both employ indeterminism to explain how such forms of causality are part of human agency.
 - 8 I should also note that the following account is compatible with accounts of systems of events that are regarded as "chaotic" or "complex" – deterministic systems that have the consequence that they are epistemically unpredictable to one extent or another though in fact they are at microscopic event levels causally deterministic as depicted by R-determinism – such as the weather. So it is certainly possible that minds/brains are chaotic in this epistemic sense – but still subject to the kind of underlying one-to-one causality that can undergird claims of incompatibilism of free will.
 - 9 As will be shown, some of these will be probabilistic accounts tending toward indeterminism.
 - 10 In this essay I will focus on the van Inwagen/Ginet accounts for M-causality and Beebe and Mele for the most prominent E-causal one.
 - 11 Laplace's demon is the real predecessor for these views (Laplace 1951).
 - 12 Though note that Campbell (2007) shows that this argument is defective in making unwarranted assumptions.
 - 13 To van Inwagen's credit he has acknowledged some of these, particularly with respect to his original assumption of inference principle Beta, though he still maintains that some version of the Consequence Argument is correct (van Inwagen 2000).
 - 14 Note that not all interpretations of QT are indeterministic, but that experiments based on Bell's inequalities certainly seem to favor indeterminism (Bell 2004). For a more nuanced view of these experiments, see Buchanan (2014).
 - 15 Beebe and Mele (2002).
 - 16 A significant review of this issue is found in Cohen and Callender (2009).
 - 17 Beebe and Mele (2002).
 - 18 This amounts to a luck problem with indeterminism that flows over into this view (Beebe and Mele 2002).
 - 19 I'm aware that Bohmian deterministic interpretations and Everett-Wheeler many-world interpretations would in different ways side-step metaphysical indeterminism. However, clearly many physicists follow interpretations of QT that are inherently metaphysically indeterministic, and that alone justifies this pragmatic line of inquiry.
 - 20 Cartwright (1983) and other of her work as well as van Fraassen (1980).

- 21 Even many libertarians admit this in effect – that there are very few instances of indeterminism of FW in life – van Inwagen, Kane, C. A. Campbell, etc. allow that exercises of libertarian free will are relatively rare. So the preponderance of evidence does not favor that view.
- 22 Davidson (2001) and Whitehead (1978) are the event ontologist inspirations here.
- 23 This is pragmatically justified by how science uses measurement.
- 24 This definition serves notice that MP as used here is always relativized to the actual world, thus snipping off wider modal worries about what *might* be metaphysically possible.
- 25 There has been recent experimental work that may well challenge these claims to mere subjective access to mental contents eventually, but it's certainly premature to declare the death of the claim to privileged access.
- 26 What this entails is that these mental event diagrams are consistent with any and all mind–body views of human nature, where the abstract space axis is needed for dualism and dual-aspect theories, but reverts to a space axis for physicalism (along with reversion to assumptions about SC in that latter case).
- 27 Again, this is only the case if dualism or dual-aspect views of human nature is true.
- 28 As this chapter goes to press, it came to my notice of the publication of Woodward 2021. This work is a thoroughgoing justification of the very kind of causality invoked by the R-determinism advocated here, and I would think abets my pragmatic posture as well.
- 29 I wish to thank my co-editors Joe Campbell and Kristin Mickelson for criticisms, suggestions, and encouragement that improved this chapter, which was 30-plus years in the making while teaching my Introduction to Philosophy course focused on the free will problem (White 1990, 1996). Of course their endorsement of my view however flawed it may be is not to be inferred.

Bibliography

- Beebe, H. and Mele, A. (2002). Humean compatibilism. *Mind* 111 (442): 201–223.
- Bell, J.S. (2004). *Speakable and Unspeakable in Quantum Mechanics*, 2e. Cambridge: Cambridge University Press.
- Bourget, D. and Chalmers, D.J. (2014). What do philosophers believe? *Philosophical Studies* 170 (3): 465–500.
- Buchanan, M. (2014). Clear as a Bell. *Nature Physics* 10: 703.
- Campbell, J.K. (2007). Free will and the necessity of the past. *Analysis* (67) 2: 105–111.
- Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Oxford University Press.
- Clarke, R., Capes, J., and Swenson, P. (2021). Incompatibilist (nondeterministic) theories of free will. *The Stanford Encyclopedia of Philosophy* (January 2022 Edition), (ed. E.N. Zalta). <https://plato.stanford.edu/archives/fall2021/entries/incompatibilism-theories>.
- Cohen, J. and Callender, C. (2009). A better best system account of lawhood. *Philosophical Studies* 145 (1): 1–34.
- Davidson, D. (2001). *Essays on Actions and Events: Philosophical Essays, Volume 1*. Oxford: Clarendon Press.
- Fenton-Glynn, L. (2017). A proposed probabilistic extension of the Halpern and Pearl definition of ‘actual cause’. *The British Journal for the Philosophy of Science* 68 (4): 1061–1124.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge, UK: Cambridge University Press.
- Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.
- Hitchcock, C. (2021). Probabilistic causation. *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), (ed. E.N. Zalta). <https://plato.stanford.edu/archives/spr2021/entries/causation-probabilistic>. (Accessed on 25th May 2023).
- Kane, R. (2004). Agency, responsibility, and indeterminism: reflections on libertarian theories of free will. In: *Freedom and Determinism* (ed. J.K. Campbell, M. O'Rourke and D. Shier), 70–88. Cambridge, MA: MIT Press.

- Laplace, P.S. (1951). *A Philosophical Essay on Probabilities*, translated into English from the original French 6th ed. by F.W. Truscott and F.L. Emory. New York: Dover Publications, 4.
- Lewis, D. (1986). Postscripts to 'Causation'. *Philosophical Papers*, Volume II, Oxford: Oxford University Press.
- Mawson, T.J. (2011). *Free Will: A Guide for the Perplexed*. New York: Continuum.
- O'Connor, T. (2009). Agent-causal power. In: *Dispositions and Causes* (ed. T. Handfield), 189–214. Oxford: Clarendon Press.
- van Fraassen, B.C. (1980). *The Scientific Image*. New York: Oxford University Press.
- van Inwagen, P. (1983). *An Essay on Free Will*. New York: Oxford University Press.
- van Inwagen, P. (2000). Free will remains a mystery. *Philosophical Perspectives* 14: 1–20.
- Vohs, K.D. and Schooler, J.W. (2008). The value of believing in free will encouraging a belief in determinism increases cheating. *Association for Psychological Science* 19 (1): 49–54.
- Watson, G. (1987). Free action and free will. *Mind* xcvi: 145–172.
- White, V.A. (1990). The single-issue introduction to philosophy. *Teaching Philosophy* March, 13–19.
- White, V.A. (1996). Single-topic introductory philosophy – an update. *Teaching Philosophy* June, 137–144.
- Whitehead, A.N. (1978). *Process and Reality*. New York: The Free Press.
- Woodward, J. (2021). *Causation with a Human Face*. New York: Oxford University Press.

(In)compatibilism

KRISTIN M. MICKELSON

1 The Problem of Determinism and Its Candidate Solutions

The terms ‘compatibilism’ and ‘incompatibilism’ were coined in the early 1960s and quickly became part of the basic vocabulary of *the classical analytic paradigm*, the dominant research paradigm of the classical period in the free-will debate (c. 1965–1985).¹ Philosophers working in the classical analytic paradigm, hereafter *classical analytic theorists*, used the phrase ‘free-will thesis’ to name the thesis that we, that is, ordinary human beings, exercise free will. The term ‘incompatibilism’ was introduced to name the view that the free-will thesis is *incompatible* with determinism, and ‘compatibilism’ was introduced to name the view that the free-will thesis is *compatible* with determinism.² Following Peter van Inwagen – who crystallized the distinctive assumptions and methods that structured the classical analytic paradigm in *An Essay on Free Will* (1983) – most philosophers still speak as if the central problem of free will and determinism is *The Compatibility Problem*, roughly the challenge of settling whether compatibilism or incompatibilism is true.

The classical analytic paradigm started to degenerate almost as soon as it formed, with the advent of “Frankfurt examples” (see Haji Ch. 6 this volume) playing a critical role in its demise. As one would expect, framing the free-will debate in the anachronistic jargon of a degenerated research paradigm is not a benign practice. In this section, I shed new light on the structure of the classical analytic Compatibility Problem by juxtaposing it with a more contemporary – and, at the same time, more traditional – way of characterizing the problem of free will and determinism and its array of candidate solutions. In the next section, I look more closely at the correlation and relevance relations which classical analytic theorists – and those working in their wake – have conflated under the label “incompatibility.” In doing so, I clear the way for contemporary free-will theorists to escape the jargon, narratives, and question-begging background assumptions of the outdated classical analytic paradigm.

A Companion to Free Will, First Edition. Edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White.
© 2023 John Wiley & Sons, Inc. Published 2023 by John Wiley & Sons, Inc.

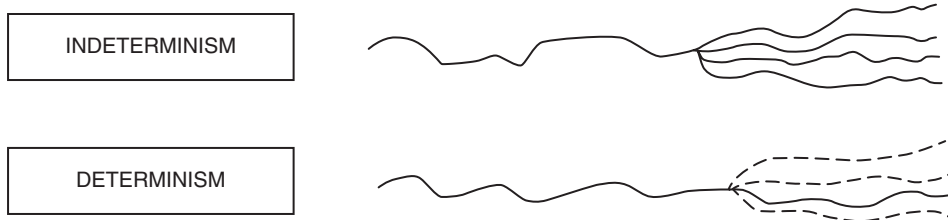
1.1 Defining Determinism

Let us start with the notion of determinism that is central to the traditional problem of free will and determinism (hereafter, *the problem of determinism*). William James, who was one of the first philosophers to use the term ‘determinism’ in the free-will literature, characterized the doctrine as follows:

What does determinism profess? It professes that those parts of the universe already laid down absolutely appoint and decree what the other parts shall be. The future has no ambiguous possibilities hidden in its womb; the part we call the present is compatible with only one totality. Any other future complement than the one fixed from eternity is impossible. The whole is in each and every part, and welds it with the rest into an absolute unity, an iron block, in which there can be no equivocation or shadow of turning. (James 1884/1897)

In this passage, James describes a universe in which one unique future is predetermined (unconditionally pre-fixed, made inevitable) by factors which were fully in place before that future unfolds.³

We can emphasize the radical nature of determinism by contrasting it with the type of “garden of forking paths” timeline that unfolds when determinism is false (i.e. when indeterminism is true) with the following pair of diagrams from Peter van Inwagen (1990):



In the top diagram, the solid lines represent that there is at least one point in this timeline relative to which there are several open “alternative futures.” A person in the top timeline would have access to multiple futures, analogous to the way in which a person has access to different routes forward when she comes to a fork in a river or road (van Inwagen 1990, p. 277). By contrast, the bottom timeline shows that when determinism is true, we never “confront a sheath of possible futures” in the actual timeline of our lives (van Inwagen 1990, p. 277); only one timeline is actually open to us, and every apparent fork in the road is an illusion.

The idea that the future is made inevitable by naturalistic forces beyond our control has been driving the free-will debate for millennia, having its roots in the naturalistic (cause-and-effect) account of fate developed by the Stoics (Bobzien 1998; Pereboom 2009, pp. 5–16; see also Kane 2002, p. 6).⁴ For present purposes, I will set aside questions about precisely which features of the world do the critical future-fixing work when determinism is true (e.g. causation, laws of nature, etc.). For simplicity, I will use the phrase “determinism-related factors” to refer to those features of the world (whatever they are) which account for the state-by-state and moment-by-moment evolution of the world when determinism is true, that is, the factors which do the “work” of making one unique future inevitable when they obtain.

In my estimation, the traditional debate over the *relationship* between free will and determinism rightly captures our attention because it forces us to confront *the problem of free will*. In short, the problem of free will is the challenge of identifying the *nature* of free will so that

we can answering the more practically pressing question of whether or not we (ordinary humans) *have* free will. When the dialectical role of determinism is seen in this way – that is, as a way of generating and evaluating candidate solutions to the problem of free will – it appears that the problem of determinism has two main components, what I call *the correlation problem* and *the explanation problem* (Mickelson 2019a, 2019b, 2021a).⁵ Let us consider these two problems in turn.

1.2 The Correlation Problem

The central challenge of the correlation problem is captured in one familiar question:

The Correlation Problem: Is it metaphysically possible for a typical human to exercise free will in a world at which determinism is true?⁶

The long history of the free-will debate demonstrates that people deeply disagree about whether the answer to this question is “yes” or “no.” Those who think that the correct answer to the correlation problem is “no” thereby embrace the view that free will – whatever else might be true of it – is not the sort of thing that could be exercised by a normal human in a world at which determinism is true. For these theorists, the intuitive “no free will” judgment elicited by thought experiments involving humans acting in a world at which determinism is true is a *data point* that must be accommodated by any viable theory of free will. That is, to answer “no” to the correlation question is to commit oneself to a desideratum on any viable theory of free will, namely: it must include at least one necessary condition which cannot be satisfied by a normal human living in a world at which determinism is true. By contrast, those who answer “yes” thereby reject that there is any such desideratum on a viable account of free will. In other words, determinism scenarios reveal that people not only have different intuitions about what free will is, but they also disagree about the standards by which a theory of free will is to be judged.

Despite the dialectical significance of the “yes” and “no” answers to the correlation problem, there is no standard jargon which unequivocally tracks the modal views associated with them. To remedy this, I have introduced new terminology (Mickelson 2012, 2015a, 2015b, 2017, 2019a, 2019b, 2021a). To give a “no” response to the correlation problem is to claim that human free will and determinism-related factors are *impossible*, that is, it is metaphysically impossible for them to co-exist. In light of this, I say that those who answer “no” to the correlation problem are proponents of an *impossibility solution* to the correlation problem, while those who answer “yes” endorse a *compossibility solution*. Correspondingly, *impossibilism* is the view that the impossibility solution to the correlation problem is correct, and *compossibilism* is the view that the compossibility solution is correct.

2 The Explanation Problem

Impossibilism answers one problem only to raise another: unlike the compossibilist, the impossibilist must explain or account for the impossibility of human freedom and determinism-related factors. To solve this *explanation problem*, the impossibilist must answer the following questions:

The Unmet Condition Challenge (E1): Which necessary condition C on free will cannot be satisfied by a normal human when determinism is true?

The Condition Underminer Challenge (E2): What prevents a normal human from satisfying condition C when determinism is true?⁷

Although the E1/E2 distinction is present in contemporary discussions of free will, free-will specialists have not made an effort to track this distinction or that between an impossibility solution to the correlation problem and extant solutions to E2 of the explanation problem. Presumably, this is at least partly because their understanding of the problem of determinism has been shaped and constrained by the distinctive jargon and narratives of the classical analytic paradigm. Given these historical considerations, let us begin by looking at the bipartite explanation problem through the narrow lens of the classical analytic paradigm.

2.1 *Classical Analytic Solutions to the Correlation and Explanation Problems*

Within the classical analytic paradigm, the questions associated with E1 and E2 were not distinguished from a “no” answer to the correlation problem, and it’s easy to understand why. Given the background assumptions of that paradigm, a complete solution to the explanation problem follows directly from an impossibility solution to the correlation problem. How so? Classical analytic theorists agreed that free will is (or requires) an *ability to do otherwise*. Since it is standard practice to attach the term ‘classical’ to view which presuppose that free will is an ability to do otherwise, it is natural to refer to the impossibilist who accepts the classical view of free will as a *classical impossibilist*. The classical impossibilist must interpret her solution to the correlation problem as providing key information about what this freedom-relevant ability to do otherwise amounts to. Specifically, the classical impossibilist commits to the view that free will, that is, the ability to do otherwise, cannot be exercised in which one unique future is made inevitable by determinism-related factors (given the facts of the past). Free will is a more robust ability to do otherwise, roughly an ability to “settle” which of the multiple futures available to her at the moment of action, in the forking-paths timeline of her life, comes to pass; in counterfactual terms, an actor exercises the relevant ability to do otherwise in performing A only when it is true of the actor that, holding fixed the laws and the facts of the past prior to the action, the actor still *could have done otherwise* than perform A, that is, she could have *avoided* or *refrained* from doing what she actually did and done something else instead (e.g. van Inwagen 1983, 2017; Balaguer Ch. 9 this volume).

Notably, classical analytic theorists also took for granted the truth of *anthropocentric possibilism*, the view that it is at least metaphysically possible for an ordinary human to exercise free will (e.g. van Inwagen 1983; Clarke 2003; Vihvelin 2013).⁸ Any account of free will that is consistent with both impossibilism and anthropocentric possibilism is now commonly classified as a *broadly libertarian account* of free will.⁹ This means that all impossibilists working in the classical analytic paradigm were committed to a broadly libertarian solution to E1.¹⁰ I will hereafter refer to the general solution to E1 endorsed by classical impossibilists, according to which some libertarian interpretation of the ability to do otherwise is correct, the *classical libertarian solution* to E1.¹¹

According to the classical libertarian solution to E1, it is metaphysically possible for a normal human to exercise free will only if there is some type of *actual-sequence leeway* in the evolution of the universe. There is actual-sequence leeway in the evolution of the universe only if there is at least one point in time at which, holding fixed the laws and facts of the past, more than one future might unfold. That is, more than one future is *accessible* from some point in the actual timeline, just as we see in van Inwagen’s “forking paths” depiction of

indeterminism (in van Inwagen's diagram seen earlier). Recall that determinism, by definition, is true only when there is literally zero actual-sequence leeway in the evolution of the physical world, that is, when exactly zero "forking paths" are permitted in the actual timeline. This means that if determinism were true, certain determinism-related factors would rule out the sort of actual sequence leeway – *whatever type that may be* – that an exercise of free will requires (given the libertarian take on the notion of an ability to do otherwise). Within the classical analytic paradigm, then, an impossibility solution to the correlation problem implies the classical libertarian solution to E1, which in turn implies the *classical incompatibilist solution* to E2: when determinism is true, it is *owing to determinism-related factors* that people cannot exercise the ability to do otherwise that an act of free will requires. For ease of reference, I will hereafter refer to this line of reasoning – from impossibilism, via the classical libertarian solution to E1, to the classical incompatibilist solution to E2 – as the *classical bridge inference*.¹²

Now that we have a better grasp on the classical analytic characterization of the problem of determinism, we can understand why classical analytic theorists believed – as their Compatibility Problem narrative tells us – that the problem admits only two (mutually exclusive and collectively exhaustive) candidate solutions. On the one hand, classical analytic theorists could adopt an impossibility solution to the correlation problem which, given the background assumptions of the classical analytic paradigm, came theoretically loaded – via the classical bridge inference – with the classical libertarian solution to E1 and the classical incompatibilist solution to E2. This complex position is now known as *classical incompatibilism*. On the other hand, the classical analytic theorists could embrace a compossibility solution to the correlation problem and proclaim that the explanation problem is a pseudo-problem generated by a false (impossibility) solution to the correlation problem and, so, all candidate solutions to the explanation problem are wrong. The latter position is now known as *classical compatibilism*. For those working within the classical analytic paradigm, then, it was quite sensible to summarize the problem of determinism as a debate about whether classical compatibilism or classical incompatibilism is true.

2.2 Contemporary Solutions to E1 of the Explanation Problem

The classical analytic paradigm has fallen out of favor, largely because many philosophers are unwilling to grant its presumptions that free will is/requires an ability to do otherwise and that anthropocentric possibilism is true.¹³ Indeed, a growing number of contemporary philosophers explicitly reject both anthropocentric possibilism and the classical definition of 'free will'. For example, Galen Strawson (1986, 1994, 2002, 2008, 2011) and Derk Pereboom (2001, 2014) are impossibilists who reject the classical libertarian solution to E1, favoring instead a type of *source solution* to E1 according to which free will requires one to be the genuine *source* of one's action. Although their accounts of freedom-relevant sourcehood are substantively different, Pereboom and G. Strawson agree that it is metaphysically impossible for a normal human to satisfy the necessary source condition on free will.¹⁴ We might say, then, that G. Strawson and Pereboom are each *source impossibilists*, insofar as each embraces a strict source solution to E1 and, moreover, each is an *anthropocentric source impossibilist*, insofar as each embraces the view that it is metaphysically impossible for an ordinary human to satisfy the source condition on free will. However, whereas G. Strawson's understanding of the source condition leads him to endorse unqualified *impossibilism*, that is, the unqualified view that it is metaphysically impossible for any being to exercise free will, Pereboom's understanding of the source condition leads him to reject impossibilism (e.g. Pereboom 2001, p. 132).

In addition, some philosophers have suggested a type of hybrid solution to E1, proposing that an ability to do otherwise is required to satisfy the source condition on free will. For example, Robert Kane's "self-forming actions" account of free will seems to be a hybrid solution of this kind (e.g. Kane 1996, 2004, 2016, 2019; see also Balaguer CH. 9 this volume). As such, it would now be question-begging to frame the free-will debate, as classical analytic theorists once did, against the background assumptions that anthropocentric possibilism is true and that some version of the classical (ability-to-do-otherwise) characterization of free will is correct.

The fact that contemporary impossibilists disagree about the nature of free will in a more fundamental way than their classical predecessors makes it easier to understand why contemporary free-will theorists struggle to provide a precise yet reasonably uncontroversial characterization of *free will* (and/or definition of 'free will') for use in their debate. The more precise one's proposed characterization, the more likely that one camp or another in the debate will find it question-begging; the more generic the proposal, the more likely that the referent of 'free will' will not be fixed securely enough to ensure that the interlocutors in the debate are disagreeing about the same thing (rather than having a mere verbal dispute in which two groups talk past each other because they are using the same phrase 'free will' to pick out different things).

There is no simple fix for this dialectical difficulty, but neo-classical and non-classical theorists have generally agreed to fix the referent of 'free will' by saying that free will is a type of *control* or *up-to-one-ness* that is necessary for moral responsibility, such that a person is morally responsible for an action A only if she exercises this type of control or exhibits this type of up-to-one-ness when performing A. Since it would be easy to find objections to any attempt to give a dialectically neutral characterization of free will (or definition of 'free will'), I will not attempt to provide one here; I leave readers to weigh the pros and cons of popular working definitions of 'free will' for themselves.

2.3 Contemporary Solutions to E2 of the Explanation Problem

Philosophers who are unaware of recent shifts away from the classical analytic paradigm may find it strange, even absurd, to suggest that there is room for a substantive and philosophically interesting debate about what *explains* and/or *accounts for* the lack of free agents in worlds at which determinism is true. Readers in this camp might wonder: Assuming that impossibilism is true, isn't it just obvious – so obvious that it may go unstated – that certain *determinism-related factors* play at least some role in making people unfree when determinism is true? Does it really matter in the end whether the impossibilists disagree about the fine-grained details regarding *which* determinism-related factors – the causation, the laws of nature, a conjunction of causal factors and facts about the past, etc. – pose a threat to human freedom, or whether determinism-related factors preclude free will *on their own* or only in conjunction with some other factors beyond our control?

Readers drawn to such questions must be reminded that the contemporary free will debate is no longer constrained by the assumptions and methods of the classical analytic paradigm. Once one abandons the classical analytic stipulation that 'free will' denotes an ability to do otherwise, the classical bridge inference is no longer a viable way to close the "explanatory gap" between an impossibility solution to the correlation problem and the classical incompatibilist solution to E2 of the explanation problem. Indeed, for all that is stated by impossibilism, it might simply be a *brute fact* that no one acts freely when determinism is true.¹⁵ Of course, few would accept that the impossibility of free will and determinism-related

factors is a brute fact; assuming impossibilism is true, there must be some *better explanation* for its truth than that. But what is that better explanation? This brings us to the central point of this section: influential impossibilists can and do disagree about what makes people unfree when determinism is true. That is, contemporary impossibilists, unlike their classical predecessors, substantively disagree about the correct solution to E2.

The suggestion that the literature already includes a variety of philosophically interesting and conflicting solutions to E2 may take some readers by surprise. After all, philosophers generally speak as though all impossibilists agree that determinism-related factors keep people from acting freely when determinism is true, but are divided about whether such factors preclude free will *because they keep people from exercising the ability to do otherwise* or, rather, such factors preclude free will *because they keep people from being the source of their own actions* (e.g. Fischer and Ravizza 1998, p. 151, Kane 2002, p. 6; Griffith 2017, p. 2; Vihvelin 2022). Unfortunately, the existence and contours of this lively in-house dispute among impossibilists have been obscured in part because the dispute cannot be adequately characterized using the jargon and narratives of the classical analytic paradigm.

For purposes of categorizing extant solutions to E2, it is helpful to distinguish between two basic categories of factors which are, arguably, *beyond one's control*. First, there are two types of actor-extrinsic factors. On the one hand, there are states of physical world outside (i.e. external to the properties which constitute) a given actor. These actor-extrinsic factors include both aspects of one's immediate environment (e.g. the physical properties of the room one is sitting in) and complete states of the world in the remote past, for example, states which were obtained before the first human was born (e.g. there were no humans when dinosaurs roamed the Earth). On the other hand, there are actor-extrinsic factors which govern or otherwise account for the way in which the universe evolves from one state to another over time. For example, *determinism-related factors* are those factors which, given the state of the physical universe at one time, ensure the fine-grained state of the physical universe at the next moment--and every moment thereafter, until the end of time. The specific nature of the factors which account for the evolution of the universe is a matter of debate, for example, they may be causal relations, laws of nature, and so on. Second, there are features of an actor which may be (arguably, at least) beyond an actor's control. These include the properties which an actor has at a given moment (e.g. one's precise brain state at a given moment), features an actor has during their entire lifespan (e.g. one's genetic endowment), as well as an actor's essential properties, if there are such things (e.g. those properties which establish an actor's cross-world identity, such that we can posit that an actor in one possible world is – or at least has a counterpart in – some other possible world).

We might introduce a completely new set of labels to track the three mutually exclusive and collectively exhaustive classes of factors described earlier, but I will instead import and refine a few terms that have become commonplace in discussions of the paradox of moral luck (e.g. Hartman Ch. 10 this volume). This choice reflects my conviction that the problem of determinism and the paradox of moral luck are best understood as two rhetorically distinct frameworks for investigating *the problem of free will*, roughly the problem of identifying what free will is and whether we (i.e. ordinary humans) have it.¹⁶ Like the problem of determinism, the paradox of moral luck is grounded partly in the observation that our normal practices of blame and punishment seemingly presuppose that we have the type of control required for basic-desert moral responsibility (McCormick Ch. 24 this volume), and yet close examination of individual actions indicates that no human has that type of control – rather, every human action is settled mostly – if not entirely – by factors beyond our control.¹⁷ Moral-luck theorists have normalized the practice of using 'luck' as shorthand for the more cumbersome phrase

“factors beyond one’s control” (Hartman 2017, pp. 23–31; Anderson 2019; Statman 2019), and I will follow that convention here. Inspired by Nagel (1976, 1979, 1986), I will speak of three basic types of luck: causal luck, circumstantial luck, and constitutive luck.¹⁸

For present purposes, let us say that a person is subject to *causal luck* when one lacks control over the factors which settle how the world evolves from one state of affairs to another (e.g. causal relations, causal laws, laws of nature, or the like). Let us say that one is subject to *circumstantial luck* when one lacks control over the non-causal/agent-extrinsic states of affairs (e.g. states of the universe prior to one’s birth). Finally, one is subject to *constitutive luck* when one lacks control over one’s own constitutive properties (e.g. one’s genetic endowments with regards to intelligence and personality, or that one is an ordinary human rather than some other type of subhuman or superhuman creature).¹⁹ Using these three categories of luck, we can easily identify a variety of extant solutions to E2: *causal luck solutions*, *circumstantial luck solutions*, *constitutive luck solutions*, and *hybrid solutions*. Let’s consider each in turn.

A strict *causal luck solution* to E2 proposes that it is impossible for people to act freely when determinism is true because there are certain nomological and/or causation-related factors which make it the case that no human can exercise free will. The classical incompatibilist (“owing to determinism”) solution to E2 may be classified as a type of causal luck solution. That said, there is plenty of room to wonder what this classical causal luck solution really amounts to. Philosophers commonly speak as though determinism per se precludes free will. Taken literally, however, this is an untenable response to E2. Why? Determinism has the ontological status of a proposition, and this proposition merely provides a *description* of a certain kind of world. There is little reason to suppose that a description (even if true) could itself pose a threat to free will; it is not the description, but *what it describes*, that has the ontological standing to preclude the exercise of free will (e.g. Hermes and Campbell 2012; Hermes 2013).

Philosophers who contend that certain evolution-governing factors *described by* determinism have the ontological standing required to undercut human freedom (as opposed to the thesis of determinism itself) carry the dialectical burden of specifying *which* evolutionary factors do the freedom-undermining work when determinism is true. Philosophers often pinpoint the *deterministic* quality of causation (laws of nature, or the like) which obtain when determinism is true as a specific threat to free will (e.g. Pereboom 2001, 2014; Sartorio 2016). Notably, though, the laws of nature may be deterministic in a perfectly standard sense of the term ‘deterministic’ even when the type of zero-leeway determinism described by James and depicted by van Inwagen (above) is false (e.g. Stone 1998; Dennett 2003; Sehon 2010; Mickelson 2012, 2019a, 2019b).²⁰ Other potentially freedom-relevant features of the laws of nature described by determinism are their *strength*, that is, that they are “strong” rather than “weak” (Perry 2004), and their *potency*, that is, that such laws are “unconditional” rather than “conditional” (Mickelson 2019a).²¹ Given these distinctions, there is ample dialectical space for a lively in-house debate among the subset of impossibilists who endorse a causal luck solution to E2 – a group we might call *causal luck impossibilists* – to disagree about which determinism-related factors (if any) are antagonistic to free will. Given that some determinism-related factors are present even when determinism is false, such disagreements may have very interesting implications for the general project of assessing which account of free will is best (Mickelson 2019a and “Hard Times for Hard Incompatibilism,” ms.).

Impossibilists who reject broadly libertarian solutions to E1 may – and arguably must²² – categorically reject causal luck solutions to E2. For example, G. Strawson is an impossibilist, and *a fortiori* an impossibilist, who rejects the classical solutions to E1 and E2. With his Basic Argument, G. Strawson argues that we (“the folk”) intuitively accept that there is an “ultimate starting point” source condition on free will, but the type of “buck stops here”

sourcehood we intuitively want – and generally take ourselves to have – could be achieved only be a *causa sui*.²³ Satisfying the sourcehood condition, in other words, would require an act of *ex nihilo* self-creation. However, the notion of self-creating *ex nihilo* implies the existence of a *nothing-self*, which is a contradiction in terms. Clearly, the notion of *ex nihilo* self-creation is incoherent *irrespective of what the world is like* (e.g. irrespective of what type of laws of nature obtain or what the remote past was like). In other words, the Basic Argument tells us that all causal luck and circumstantial luck solutions to E2 are false; it is specifically *constitutive luck*, and not any and every sort of luck around, that “swallows everything” (Mickelson 2019b). Among other things, this means that philosophers who are still working on the outdated classical analytic assumption that a mere impossibility solution to the correlation problem guarantees that a causal luck (“incompatibilist”) solution to E2 – or, equivalently, that any argument for impossibilism is an argument for incompatibilism (a.k.a. causal luck impossibilism) – is guilty of begging the question against someone (we might call them a *constitutive luck impossibilist*) such as G. Strawson.

There seems to be little sympathy in the literature for a *strict circumstantial luck solution* to E2, that is, a solution which says that it is impossible to satisfy the condition on free will named in E1 when determinism is true strictly because people lack control over certain (non-causal, agent-extrinsic) states of affairs which obtained prior to their actions.²⁴ However, philosophers often combine circumstantial luck with another type of luck to create a hybrid solution to E2. For example, some philosophers promote a *causal/circumstantial luck hybrid solution* according to which it is the combination of our lack of control over the laws and circumstances in the past which makes us unfree. For example, Pereboom appears to be in this camp, as he contends that normal humans would lack free will when determinism is true because in such conditions one’s actions are deterministically caused *by factors beyond their control*, where it appears that the additional “factors” he is alluding to are certain past states of the world over which the actors exercised no control (Pereboom 2001, 2014). Constitutive luck solutions to E2 may also come in hybrid form, and some philosophers have suggested that circumstantial/constitutive luck hybrid solutions are more compelling than any type of causal luck solution (e.g. Nagel 1986, pp. 113–114; Latus 2001; Levy 2011, p. 96).²⁵

Summing up, we have seen that classical analytic theorists conceived of the free-will debate as an in-house dispute among anthropocentric possibilists over the correct understanding of *the ability to do otherwise*, with the classical compossibilists adopting one view and classical impossibilists adopting another. The classical libertarian solution to E1 fit perfectly with the classical incompatibilist solution to E2, which accounts for the fact that all classical impossibilists were also classical incompatibilists. Since all classical compossibilists rejected both impossibilism and the classical incompatibilist solution to E2, it made sense for classical analytic theorists to propose that the fundamental divide in the free will debate is that between (classical compossibilist) compatibilists and (classical broadly libertarian) incompatibilists. However, the free will debate has outgrown the bipartite compatibilism/incompatibilism taxonomy that we inherited from the classical analytic theorists. In the contemporary debate, philosophers are invited to select from a much wider range of fundamentally different positions.

2.4 Testing the Quality of Candidate Solutions to the Explanation Problem

The observation that impossibilists can and do disagree about the correct solution to explanation problem sheds new light on the overall dialectic of the free-will debate. Teasing apart E1 and E2 makes it easier to see that an impossibilist who aims to solve the

explanation problem must provide a solution to E2 that fits with his proposed solution to E1. By “fit,” I mean that whatever is singled out as the “threat” to free will in one’s solution to E2 must be the kind of thing that could keep a person from satisfying the necessary condition on free will named in one’s solution to E1. For example, we have seen that the internal coherence of the classical incompatibilist solution to E2 of the explanation problem was assured, via the classical bridge inference, by the classical libertarian solution to E1. Likewise, G. Strawson’s suggestion that people have an incoherent “ultimate starting point” conception of sourcehood fits well with the view that constitutive luck (i.e. the lack of relevant control over one’s constitutive properties) and not contingent determinism-related factors is what makes it impossible for a person to satisfy the source condition on free will when determinism is true.²⁶

Stepping back, the fact that impossibilists disagree about what free will is and what undermines it implies that impossibilists also disagree about what impossibilist intuitions are tracking. We must resist the temptation (encouraged by the lingering influence of classical analytic narratives) to assume that impossibilist intuitions are rational insofar as they track the freedom-undermining effects of determinism-related factors, for some impossibilists contend that they track the threat posed by the toxic combination of determinism-related factors and circumstantial factors, others contend that these intuitions track the menacing nature of constitutive luck alone, and the list goes on. These rival accounts of what makes impossibilist intuitions rational and impossibilism true cannot all be right, but they may all be wrong – after all, impossibilism may be false and all impossibilist intuitions may be faulty.

It should now be clear that embracing impossibilism is just a starting point for a person with impossibilist intuitions. His project is not done until he tells his rivals – possibilists, agnostics/neutral-inquirers, and rival impossibilists alike – a plausible and detailed story about what free will is (i.e. by answering E1) and precisely what undermines it when determinism is true (i.e. by answering E2). Indeed, there seems no way to assess – therefore no reason to *accept* – that impossibilism is true and/or that impossibilists intuitions (as opposed to possibilist intuitions) are truth-tracking until impossibilism is coupled with a specific solution to the explanation problem. Put another way, impossibilism is a useful rhetorical landmark in a complicated dialectic, but it is not among the viable endgame positions in the free-will debate.²⁷

3 The “Incompatibility” Relations of Incompatibilism

Since the early stages of the classical analytic period in the free-will debate, free-will theorists have been conflating relations which are as fundamentally different in kind as correlation and causation under the term ‘incompatibility.’ Just as there is no single incompatibility relation picked out by the term ‘incompatibility,’ the term ‘compatibility’ has become an umbrella term for a variety of non-incompatibility relations. While the failure to track these distinctions was arguably little more than a bit of bad housekeeping for those working in the classical analytic paradigm, these distinctions are indispensable to those working with a more contemporary characterization of the debate. In this section, I review the relations currently conflated under the label ‘incompatibility,’ and thereby expose the substantive ambiguity of the terms ‘incompatibilism’ and ‘compatibilism.’

3.1 Relevance Relations versus Correlation Relations

Outside of academic philosophy, people speak of an “incompatibility” relation between things when they wish to indicate that those things cannot co-exist in harmony in virtue of a *conflict*

rooted in their respective natures. So understood, incompatibility is a type of antagonistic *relevance relation*. To say that a relevance relation holds between two relata is to say that the relata are connected in such a way that the existence and/or specific properties of one relatum (e.g. its causal properties, its truth-value) are *relevant to* the existence of and/or specific properties of the other. As I am characterizing it, the class of relevance relations is very broad and includes causal relations (assuming a strong/non-Humean account of causation), settling relations, fixing relations, non-causal determining relations, grounding relations, backing relations, and explanatory “because” relationships.²⁸

Relevance relations may be fruitfully contrasted with another extremely broad class of association relations, namely *correlation relations*. As suggested by the etymology of the term ‘correlation’ – it is derived from a combination of the Latin *cor-* (“together”) and *relatio* (“relation”) – a correlation is a type of mutual relationship or connection between two or more things which may or may not be underwritten by a relevance relation. That is, unlike a relevance relation, the type of co-occurrence or co-variation associated with correlation – no matter how regular or useful for making predictions – may be a complete fluke. While the strength of a correlation is often described in statistical terms, the generic class of correlation relations is quite diverse. For example, the non-explanatory covariance relation of *supervenience* is a mere correlation relation which is sometimes contrasted with the explanatory relevance relation of *superdupervenience* (e.g. McPherson 2021). When two things stand in a relevance relation, there will also be a correlation relation between them; however, a correlation relation may or may not be underwritten by a relevance relation. Because of this asymmetry, it is always a mistake to infer that a relevance relation holds between two things on the meager grounds that they are correlated. As such, the distinction between correlation relations and relevance relations allows us to understand the *cum hoc, ergo propter hoc* fallacy in its broadest terms. We will return to the topic of correlation relations in a moment.

3.2 Three Relevance Relations: Conceptual, Metaphysical, and Logical Incompatibility

In standard discourse, *incompatibility* is an antagonistic relevance relation, that is, it is not a perhaps spurious correlation relation. Familiar examples of incompatibility, in the standard relevance-relation sense, include: *incompatible medications* (which chemically interact in a way that reduces the wanted medicinal effects or brings about negative effects for the patient), *incompatible roommates* (who have conflicting personalities), and *genetic incompatibility* (which indicates that the genetic properties of two individuals make it impossible for them to produce viable – live, non-sterile, etc. – offspring). Based on the way free-will specialists speak – that determinism *threatens* (Baker 2006, p. 313), *conflicts with* (Kane 1999, p. 218; Chisholm 1964/2009, p. 24; Ayer 1954/2009, p. 15), *rules out* (McKenna and Pereboom 2016, p. 169), *undermines* (Nahmias et al. 2006, p. 40), *is at odds with* (Pereboom 2005, p. 240), *precludes* (Mele 2006, p. 189), *vitiates* (Fischer 1994, p. 159), *is menacing to* (Haji and Cuypers 2006), *is a problem for* (Campbell 2011, p. 23), and *cannot be reconciled with* (Dennett and Caruso 2021, p. 5) free will and/or moral responsibility²⁹ – the non-specialist might reasonably suppose that incompatibilism is named after some variation of the standard antagonistic, relevance-relation notion of incompatibility. On this interpretation, ‘incompatibilism’ names the view that determinism-related factors are incompatible with – in the sense that they destroy, rule out, preclude, undermine, make impossible, conflict with – free will, that is, it names a broadly causal luck solution to E2 of the explanation problem. Notably, most philosophers consider this type of relevance-relation incompatibility to be at the heart of the “traditional” incompatibilist position (e.g. Pereboom 2001, pp. 128–129; Vihvelin 2008, 2013; McKenna 2010, p. 432; Levy 2011, p. 1, n.1; McKenna and Pereboom 2016, p. 151).

Incompatibility relations come in a variety of species, some of which may be distinguished by the ontological status of the relata. For example, when philosophers talk about the *conceptual incompatibility* of free will and determinism, they are talking about an incompatibility relation that holds between *concepts* or the *conceptual content* of certain statements (e.g. van Inwagen 1983, pp. 1, 66, 87; Nahmias et al. 2006, p. 30; Mele 2012, p. 434). In these discussions, context clearly indicates that the standard relevance-relation notion of incompatibility is at play. For example, the concepts BACHELOR and MARRIED are conceptually incompatible in the relevant sense. Unpacking the content of BACHELOR, we find that a necessary condition on being a bachelor is *being unmarried*, and this reveals a direct tension in the content of the concepts BACHELOR and MARRIED – a tension which accounts for (explains, grounds, or the like) the fact that it is impossible for anything to *be* a married bachelor. Likewise, to say that the concept of FREE WILL is incompatible with DETERMINISM is to say that there is a conflict rooted in the content of these two concepts from which it may be derived that it is impossible to exercise free will when determinism is true. It is unsurprising that philosophers working in the wake of the classical analytic paradigm would focus on the issue of conceptual incompatibility, for – as the discussion of the classical bridge inference in the previous section makes plain – the classical impossibilists’ proposal that some broadly libertarian interpretation of the ability to do otherwise is a necessary condition on free will commits them to the conceptual incompatibility of FREE WILL and DETERMINISM.

Although conceptual incompatibility is an interesting relation, it is arguably not a basic one. We can unpack this relation in either logical or metaphysical terms. In logical terms, we might flesh out the conceptual incompatibility of BACHELOR and MARRIED by saying that the proposition *that S is a bachelor* and the proposition *that S is married* are *logically incompatible*: the truth of the former proposition (in virtue of its meaning) would preclude the truth of the latter (in virtue of its meaning), and vice versa. Put another way, the statement “S is a married bachelor” is self-contradictory given the meaning of the terms in the statement. For ease of reference, I will call a logical incompatibility relation that holds necessarily a *strict logical incompatibility relation* (since the term ‘strict’ is often used in logical contexts to connote necessity).³⁰

Alternatively, we might cash out conceptual incompatibility as a *metaphysical incompatibility* relation. Continuing with our example, a thing cannot “fall under” the concepts BACHELOR and MARRIED simultaneously: any person who satisfies the necessary conditions on being a bachelor is someone who has the property of being unmarried, and it is impossible for someone who has the property of being unmarried to simultaneously have the property of being married. By contrast, BACHELOR and HAPPY are not conceptually incompatible, which we can unpack by saying that it is metaphysically possible for a single thing to exist which simultaneously “falls under” or “answers to” both concepts.

3.3 Two Correlation Relations: Inconsistency and Impossibility

The three incompatibility relations discussed so far belong to the class of non-spurious *relevance relations*, a class I have contrasted with perhaps spurious *correlation relations*.³¹ With this distinction drawn, let us further flesh out the distinction by considering different kinds of correlation relations which are routinely picked out by the terms ‘incompatible’ and ‘incompatibility’ in the contemporary free-will literature: *metaphysical impossibility* and *strict logical inconsistency*.

Metaphysical impossibility is a type of correlation relation and, as with any other correlation relation, it may be *spurious*. Two subtypes of spurious correlation are worth noting,

what I will call *indirect correlation* and *trivial correlation*. A correlation relation is *indirectly spurious* when the correlation between phenomena is brought about or otherwise ensured, via independent routes, by some third variable (a.k.a. a confounding variable). A stock example of an indirectly spurious correlation relation is the strong correlation between ice cream sales and murder rates: as ice cream sales rise and fall in a given region, so do the murder rates there. However, this is not because eating ice cream *causes* people to commit murder or that murdering inclines people toward eating ice cream. Rather, the rise in ice cream sales and murder rates is only *indirectly related* by a third variable which brings about each (via independent causal mechanisms), namely: the rise in outdoor temperatures. On the other hand, trivially spurious correlations – such as that between the number of people who drown by falling into pools each year during a given period and the number of films that Nicolas Cage appeared in each year during that period – are mere coincidence; there is no hidden third variable or other underlying story which accounts for and/or explains them.

Spurious metaphysical impossibility also comes in both indirect and trivial forms. A salient example of *indirect impossibility* is evident in the following view about the relationship between God's foreknowledge and human free will. Assume for a moment that determinism is true and that free will is metaphysically incompatible with – directly destroyed by, precluded by, undermined by – the type of causal laws described by determinism (i.e. grant that some type of causal luck solution to E2 is correct). Assume also that God exists. From these assumptions, one could plausibly argue that God knows (given his knowledge of the laws of nature and facts about the early state of the universe in which we live) what will unfold in the future timeline of this universe, including every action a given person performs. However, it does not follow from this that God's foreknowledge is a freedom-undermining factor of the world. One can reasonably hold that God's foreknowledge is *impossible* with free will, but only spuriously so. For example, one may identify some third variable, such as the causal relations described by determinism, as the feature of the world which directly destroys free will and (independently) accounts for God's foreknowledge. Here, God's foreknowledge – even if we grant that it is *a factor beyond our control* – is not metaphysically/explanatorily relevant vis-à-vis the fact that no one acts freely when determinism is true, just as the rise in murder rates is not causally/explanatorily relevant to the rise of murder rates in the example earlier.³² In short, one may plausibly affirm that God's foreknowledge is (spuriously) impossible with human free will while denying that these phenomena are metaphysically incompatible.

Metaphysical impossibility may also be completely trivial. A spurious impossibility relation is *trivial* when it holds between two phenomena which bear no direct nor indirect relevance relationship to one another. For example, fluffy kittens and round squares are impossible, but only trivially so: the existence of fluffy kittens does not keep round squares from existing (or vice versa). In this case, the impossibility follows trivially from the metaphysical impossibility of round squares alone. Although free-will experts have made no consistent effort to track the distinction between metaphysical incompatibility and spurious impossibility in recent years, the value of the distinction should be plain enough to open-minded inquirers. For example, only someone with a good grasp of this distinction is in position to grasp the upshot of G. Strawson's Basic Argument. As noted in the previous section, the Basic Argument aims to show that free will is metaphysically impossible because people's actions are settled by *constitutive* properties of the agent that are beyond that agent's control, that is, free will is impossible due to constitutive luck. It follows *a fortiori* from the impossibilist conclusion of the Basic Argument that free will is trivially impossible with all determinism-related factors (e.g. deterministic laws of nature) and, in the same uninteresting way, that

free will is trivially impossible with fluffy kittens, dirty diapers, and lilac bushes in bloom. So, if the Basic Argument is sound, then determinism-related factors – despite being among the factors beyond human control – do not even partly account for or explain the lack of free agents in worlds at which determinism is true; determinism-related factors are completely irrelevant to – and, hence, are not metaphysically incompatible with – free will. In sum, the Basic Argument promotes a solution to E2 which implies that causal luck solutions (including the classical incompatibilist solution) to E2 are categorically false – for this argument concludes that determinism-related factors are *trivially spuriously* impossible with free will, which is something that all causal luck solutions to E2 deny (since the latter, by definition, propose that a *non-spurious* antagonistic relevance relation holds between these relata).

Corresponding to the sharp distinction between the metaphysical relations of metaphysical impossibility and metaphysical incompatibility, we may draw a distinction between the logical relevance relation of *strict logical incompatibility* and its correlation counterpart *strict logical inconsistency* (a.k.a. *logical non-compossibility*, *logical impossibility*).³³ Here, I assume a common non-relevance definition of ‘logical inconsistency’: Statements are *logically inconsistent* if and only if they cannot have the truth-value *true* at the same time.³⁴ In the language of first-order propositional logic, we may say that propositions p and q are logically inconsistent if and only if $\sim(p \cdot q)$ is true, or, equivalently, that p materially implies the negation of q , that is, that $(p \rightarrow \sim q)$.³⁵ The latter way of expressing the logical inconsistency relation is particularly useful for our purposes since it is uncontroversial and widely known that material implication is *not* a relevance relation (e.g. Mares 2020). In short, to say that p and q are logically inconsistent is to give us information about how the truth-values of p and q covary; it tells us nothing about *why* the truth-values covary in this way. For example, such claims do not tell us that the inconsistency is due to some kind of *conflict* between propositions p and q , for example, that their inconsistency is due to an underlying (syntactic and/or semantic) *incompatibility*.

Like metaphysical impossibility, a logical inconsistency relation may or may not be spurious. For example, given the spurious correlation between ice cream sales and homicide rates (noted earlier), the material conditional “If the rate of ice cream sales rises during period P, then the rate of homicides increases during period P” is true, but it would be a mistake to think that this material conditional claims that the rise of ice cream sales is directly relevant to (accounts for, causes, etc.) the rise in the number of homicides during P. Moreover, an inconsistency relation between two statements (and/or propositions) may be spurious even when it holds necessarily, that is, when it holds in all possible worlds; in such cases, let us say that a *strict logical inconsistency* (or simply *strict inconsistency*) obtains. For example, consider the following material conditional: “If Joe has a picture of a round square, then $2 + 2 = 5$.” This strange conditional is not only true, but necessarily true: a false antecedent suffices to make a material conditional true, and the antecedent of this conditional is not just false, it is necessarily false. Here, then, we have an example of a material conditional that is necessarily true even though there is no relevance relationship whatsoever (either in syntax or semantics) between the propositions expressed in the antecedent and consequent. Indeed, the statement “Joe has a picture of a round square” is strictly, though trivially, inconsistent with every possible statement.

By contrast, a strict logical incompatibility relation holds only when there is a direct conflict between two (or more) strictly inconsistent statements.³⁶ For example, the proposition *that John is a bachelor* is strictly logically incompatible with the proposition *that John is married*, for the term ‘bachelor’ means (among other things) unmarried. By contrast, let us say that *the round-square thesis* states that a round square exists and *the fluffy-kitten thesis* states that

a fluffy kitten exists. The fluffy-kitten thesis is strictly inconsistent with the round-square thesis, but the inconsistency is spurious; the two theses are not strictly (or otherwise) logically incompatible. To put the point another way, the material conditional “If the fluffy-kitten thesis is true, then the round-square thesis is false” is true – indeed, necessarily true – but only trivially so.

3.4 Characterizing Compatibility

For each of the relations discussed above, questions arise about what we should say when we wish to *deny* that two things stand in that relation. For example, any two things which are not impossible are *compossible*, that is, their co-existence is metaphysically possible. As such, for any two phenomena A and B we might select, the claim “A is impossible with B” will be semantically equivalent to the claim “It is not the case that A is compossible with B.” Likewise, any two propositions which are not inconsistent are *consistent*. Unfortunately, matters are more complicated when it comes to denying that a relevance-relation of incompatibility (whether metaphysical, logical, or conceptual) obtains. Any two things which are metaphysically incompatible are also impossible, but not vice versa (since impossibility may be spurious but metaphysical incompatibility cannot be). As such, one may deny that two things are metaphysically incompatible without denying that the things are impossible (and the same goes, *mutatis mutandis*, for logical incompatibility and logical inconsistency). Now, we could agree to use ‘compatibility’ very narrowly, such that two things are *compatible* so long as they are not *incompatible*. However, if philosophers were to use ‘compatibility’ in this narrow way while holding fixed the wording in standard definitions of ‘compatibilism,’ anyone who categorically rejects causal luck solutions to E2 would qualify as a compatibilist. In other words, this use of ‘compatibilism’ would open up logical space for compatibilists who are impossibilists (Mickelson 2012, 2015a) – and, indeed, this space has been claimed by some impossibilists who, by all appearances, choose to identify as “compatibilists” in order to emphasize that they reject incompatibilist (i.e. causal-luck) solutions to E2 of the explanation problem (e.g. Levy 2011, p. 1; see also Mickelson 2019b).

4 Conclusion

The terms ‘incompatibilism’ and ‘compatibilism’ were developed for use within the classical analytic paradigm, and they have become substantively ambiguous since free-will theorists started using them outside that dialectical context. It is no longer the case that there are only two basic positions one may take regarding the relationship between free will and determinism, and it may be best for us to stop using language which misleadingly implies otherwise. To repair the discourse, we might introduce new terms which reflect the distinct relations and explanations that are most important in contemporary discussions of the problem of determinism. As suggested previously, we might avoid the term ‘incompatibilism’ by referring to philosophers who accept an impossibility solution to the correlation problem and a causal luck solution to E2 as *causal luck impossibilists* (as opposed to referring to them with the degenerated term ‘incompatibilists’), and to causal luck impossibilists who also endorse a source solution to E1 as *causal-luck source impossibilists*. By contrast, we might classify impossibilists who endorse a source solution to E1 and a constitutive luck solution to E2 as *constitutive luck source impossibilists*, and so on. While I find questions about how to best define and/or redefine certain basic terms of art to be worthy of serious consideration

(e.g. Mickelson 2015a), such matters cannot be settled here. The modest goal of the present chapter has been to clarify the language and dialectical structure of the contemporary debate in a way that leaves the reader with a better understanding of the problem of determinism and its surprising array of rival solutions.

Acknowledgments

Thanks to the University of Gothenburg and the Swedish Research Council for funding this chapter. For their help in developing the ideas presented here and/or comments on early drafts of this chapter, I am deeply indebted to Derk Pereboom, Christian Munthe, Michael McKenna, Robert Hartman, Gunnar Björnsson, Keith Lehrer, Peter van Inwagen, Seth Shabo, Randy Clarke, Kadri Vihvelin, Neil Levy, Maria Sekatskaya, Audun Bengtson, Ragnar Francén Olinder, John Eriksson, Saul Smilansky, Alex Skiles, Anna Sophia Maurin, Christian Lee, Per Milam, Sophia Jeppsson, Matthew Jernberg, Taylor Cyr, Graeme Forbes, V. Alan White, and Joe Campbell.

Notes

- 1 Peter van Inwagen isolates 1965–1985 as the “classical period” of analytic work on free will; he has also (independently) suggested that the leading free-will theorists in this period were working within a distinctive paradigm (van Inwagen 2017). The terms ‘compatibilism’ and ‘incompatibilism’ were coined by Keith Lehrer during this period. To my knowledge, the terms ‘compatibilist’ and ‘incompatibilist’ were first used in print in Lehrer’s (1960) dissertation, though more standard characterizations of these terms first appeared in print eight years later (Cornman and Lehrer 1968, p. 130). It appears that the corresponding terms ‘compatibilism’ and ‘incompatibilism’ were first used in print by van Inwagen in his (1969) dissertation, but van Inwagen credits Lehrer with the coining of these terms as well (in correspondence; see also van Inwagen 2017; 1999, p. 342, n. 2). Lehrer (in correspondence, 2020) has confirmed that he introduced ‘incompatibilism’ to name a view about the *relevance* of determinism to free will. In saying this, Lehrer means to draw on the familiar conversational notion of *relevance* (as opposed, for example, to some technical notion developed within so-called *relevance logics*), and he is clear that the relevance relation he has in mind is not fully captured by the non-relevance relation of (strict) logical inconsistency. The general notions of *strict logical incompatibility* and *conceptual incompatibility* (Sec. 2) seem to fit well with the notion of relevance that Lehrer has in mind. (The terms ‘compossibilism’ and ‘incompossibilism’ were introduced in Mickelson 2012).
- 2 The classical bipartite (in)compatibilism taxonomy of free-will views was preceded by a pre-classical tripartite taxonomy of *libertarianism*, *soft determinism*, and *hard determinism*. Hard determinism is the explanatory view that determinism is true and we do not have free will because determinism-related factors preclude free will (e.g. Pereboom 2001, p. 323). Since *free-will denialism* is the non-explanatory view that no normal human in the real world has free will, all hard determinists are denialists. *Libertarians* hold that we have free will (i.e. free-will denialism is false) and, since determinism-related factors destroy free will, it must be that determinism is false. *Soft determinists* reject free-will denialism (like the libertarians) but accept that determinism is true (like the hard determinists) – and, by implication, reject the principled view (endorsed by both hard determinists and libertarians) that determinism-related factors stand in an antagonistic relationship to free will. Classical analytic theorists shifted the focus of the free-will debate to this disagreement about the in-principle relationship between free will and determinism, thereby allowing the debate about the *nature* of free will to proceed without anyone having to take a stand on the truth-value of determinism (at the actual world). In short, each view in the pre-classical tripartite taxonomy took a

- stand on three contentious issues: (1) the existence of free will, that is, whether free-will denialism is true, (2) the truth-value of determinism, and (3) whether determinism conflicts with (undermines, precludes, or stands in some broadly antagonistic relation to) free will. Classical analytic theorists aimed to track the third issue with their Compatibility Problem.
- 3 Notice that James manages to express the idea that naturalistic elements which account for the evolution of the universe make one unique future *inevitable* (by eliminating literally all actual-sequence leeway from the world) without explicit appeals to causation or laws of nature. As such, James shows that it is possible to capture the traditional “literally zero leeway” notion of determinism without having to set foot in the “morass” of causation (see also n. 21 *infra*).
 - 4 Some philosophers assume that ‘determinism’ should be used to pick out the doctrine that *deterministic* causation and/or *deterministic* laws of nature obtain, where the term ‘deterministic’ may be defined however one likes. As a result, ‘determinism’ is now a substantively ambiguous term. For example, some philosophers hold that ‘determinism’ should be used to name the view that naturalistic factors in the physical universe eliminate all actual-sequence leeway from its evolution, but only *ceteris paribus* (such that the world has, *ceteris paribus*, one unique future) (e.g. Dennett 2003; Sehon 2010). More specifically, one might use ‘determinism’ to name the thesis that there is no actual-sequence leeway in the world *on the condition that the universe remains causally closed* (e.g. so long as the world is not prematurely destroyed in a collision with another universe in the multiverse or miraculously intervened in by God). The addition of *ceteris paribus* clauses to the Jamesian notion of determinism might seem like an improvement for those who think that free-will theorists should define ‘determinism’ in a way that reflects the best physics and/or metaphysical theories of the day. In my assessment, the problem of determinism (understood as a means of addressing the problem of free will via the correlation and explanation problems) is best framed using a doctrine which is the limiting case for minimal actual-sequence leeway (i.e. a doctrine which asserts that there is *literally zero* such leeway), for this is the most straightforward way to test whether or not the presence/absence of actual-sequence leeway is *relevant* to such things as “the ability to do otherwise.” To repair the discourse, we might introduce a new term (preferably one unrelated to any of the vocabulary terms used by the physicists) such as “inevitabilism” to name the dialectically important doctrine that one unique future is *literally* – and not simply *ceteris paribus* – inevitable given naturalistic factors which hold in the present/past (for discussion, see Mickelson 2019a and Mickelson “Hard Times for Hard Incompatibilism,” ms.; see also n. 21 *infra*).
 - 5 The general framework outlined here is not an attempt to persuade readers of what the dialectical structure of the free will debate *should* look like; it aims, more centrally, to be (loosely speaking) a rational reconstruction of what the extant free will debate *does* look like. I believe this framework offers a superior map of the territory on which the free will debate is – and traditionally has been – playing out.
 - 6 This question is typically raised after providing a vignette in which normal humans are performing actions in a world at which determinism is (by stipulation) true. In recent years, those who advocate a “no” answer to the correlation problem have begun using so-called *manipulation arguments* (Mickelson 2017) as intuition pumps for the “no” answer. For example, Alfred Mele’s *revised* Zygote Argument (introduced in Mele 2013; see also Mele 2017, 2018, 2019b, 2022) is a (relatively weak) intuition pump for an impossibilist solution to the correlation problem; unlike the *original* Zygote Argument (Mele 2006, 2008), the revised Zygote Argument does not aim to solve the explanation problem (Mickelson 2015b, 2017, 2019b). Other manipulation arguments, including Pereboom’s Four-Case Argument (Pereboom 2001, 2014) and my Master Manipulation Argument (Mickelson 2019b), speak to both the correlation problem and the explanation problem.
 - 7 It has been suggested that one answer E2 by appeal to intuition, for example, by reporting that one has the *intuition* that determinism-related factors undermine free will. I do not find this suggestion compelling. Rather, I think that one’s intuitions may lead them to form a *hypothesis* about what precludes free will, but any such hypothesis may be subject to testing which demonstrates that it is mistaken. This is perhaps most evident in the literature on manipulation arguments, which exploit the intuition that a person who is subject to manipulation is unfree even though it is essential

to the success of the argument that, contrary to initial appearances, there is nothing about the manipulation per se that poses a threat to free will. Likewise, abductive manipulation arguments may be developed to test and attack the hypothesis that determinism poses a threat to free will, placing the burden of proof on libertarian theorists to defend that hypothesis against known rivals (Mickelson 2015b, 2017, 2019a, 2019b, 2021a and “Hard Times for Hard Incompatibilism,” ms.).

- 8 The assumption of anthropocentric possibilism is evident, for example, in van Inwagen’s “mysterianism” position: he contends that free will is a mystery because he believes that we have free will even though he grants that there is a strong and unanswered case for *anthropocentric impossibilism*, that is, the view that it is metaphysically impossible for a normal human to exercise free will (e.g. van Inwagen 2000; Campbell and Lota CH. 8 this volume). Notably, it has recently become popular to use the term ‘impossibilism’ exclusively to name the maximally bold impossibilist thesis that it is metaphysically impossible for anyone (i.e. for any metaphysically possible being, however godlike) to exercise free will, and ‘possibilism’ to pick out the negation of this unqualified impossibilism (e.g. McKenna and Pereboom 2016). I follow the latter convention here, but I warn against conflating impossibilism and species anthropocentric impossibilism. While all arguments for impossibilism are arguments for anthropocentric impossibilism, the converse is not true. For example, Pereboom’s arguments for hard incompatibilism imply that anthropocentric impossibilism is true but that impossibilism is false (e.g. Pereboom 2014, pp. 4–6). By contrast, some arguments (e.g. the Basic Argument and some arguments based on the paradox of moral luck) conclude to impossibilism and, *a fortiori*, support anthropocentric impossibilism, but do this by reasoning which implies that hard incompatibilism is false (e.g. Mickelson 2019a, 2019b, and “Hard Times for Hard Incompatibilism,” ms.).
- 9 The classical libertarian solution to E1 narrowly proposes that anthropocentric possibilism is true. However, “broadly libertarian account of free will” is aptly applied to any account of free will which proposes that both impossibilism and unqualified *possibilism* (the view that it is metaphysically possible for *someone* to exercise free will) are true.
- 10 One popular test for the adequacy of libertarian solutions to the explanation problem is the so-called “Problem of Enhanced Control,” a problem which is typically aimed at *event-causal* (as opposed to *agent-causal*) *libertarianism* (e.g. Franklin 2011). Viewed through the correlation/explanation framework, the problem is roughly this: to be viable, a broadly libertarian response to E1 must identify a necessary condition on free will which cannot be satisfied when determinism is true (i.e. some necessary condition that is missing from all compossibilist accounts of free will), such that it is plausible to think that someone who satisfies that condition would (in virtue of satisfying that condition) have *more control* than would a person who satisfies the necessary conditions on free will named in the best compossibilist accounts of free will. It would be unsatisfactory for aspiring libertarians to respond to E1 by insisting that *indeterminism itself* (the mere negation of determinism) or some more specific indeterminism-related factor (e.g. *probabilistic* causation) is the necessary condition on free will that goes unmet when determinism is true. As such, those who believe that (some species of) indeterminism “helps” a person to have more control than they would have in a world at which determinism is true – namely, those who endorse a libertarian account of free will – owe their audience a story about *how* indeterminism helps. E1 captures and illuminates the dialectical demand for that story.
- 11 Arguably, the classical libertarian solution to E1 is slightly more narrow than this, since it seems to be that classical analytic theorists agreed that free will requires some degree of *contrastive control* over the future (allowing a person to settle which action is ultimately performed and which are not). Some libertarian accounts of the ability to do otherwise are less demanding, such as Robert Kane’s “self-forming actions” account (e.g. Kane 1996, 2009, 2016).
- 12 Recent debates about the Consequence Argument raise the question of what this argument is an argument for (e.g. Campbell 2007, 2008, 2010; Shabo 2011; Bailey 2013; Sartorio 2016; Capes 2019). These discussions do not take into account that there are two substantively different ways of interpreting van Inwagen’s formal statements of the Consequence Argument. First, we may interpret these logic-text proofs – which van Inwagen characterizes as the same argu-

ment done three ways – against the background assumptions of the classical analytic paradigm. In this case, van Inwagen’s formal arguments technically conclude to a conditional thesis which is equivalent to classical impossibilism, but this conclusion entails (via the classical bridge inference) classical incompatibilism. Second, we may untether these formal statements of the Consequence Argument from the background assumptions of the classical analytic paradigm (i.e. the now question-begging classical assumptions which underwrite the classical bridge inference). In this case, the conditional conclusion of van Inwagen’s arguments do not entail or suggest any particular solution to E1 or E2 of the explanation problem. When a logic-text statement of the Consequence Argument is removed from its original dialectical context, its conditional conclusion must be taken as it stands, that is, as an inconsistency thesis which provides nothing more than a solution to the correlation problem. This non-classical Consequence Argument – just like Mele’s non-classical Zygote Argument (Mickelson 2015b, 2021a, and “Motte-and-Bailey Incompatibilism,” ms.) – may be used to motivate a constitutive luck solution to E2 of the explanation problem (which implies that all causal luck solutions are wrong). For further discussion, see my “The Consequence Argument: An Argument for Incompatibilism?,” ms.).

- 13 Frustration with the community’s commitment to the classical analytic definition of ‘free will’ played a role in John Martin Fischer and Mark Ravizza’s development of *semi-compatibilism*, a view about the compossibility of determinism-related factors and the type of control over/ownership of one’s actions that is required for *moral responsibility* (e.g. Fischer and Ravizza 1998). According to the semi-compatibilist, determinism-related factors pose no threat to moral responsibility irrespective of whether such factors destroy our ability to otherwise (where this is understood neutrally between the compossibilist and impossibilist reading). Notably, the latter claim is also accepted by some source compossibilists (e.g. Harry Frankfurt, as well as some anti-compossibilist source impossibilists, for example, G. Strawson, as discussed in main text). When we hold fixed the anachronistic classical definition of ‘free will,’ so-called semi-compatibilists are committed to the odd-sounding claim that moral responsibility does not require free will.
- 14 G. Strawson argues that the source condition on free will is impossible for anyone or anything to satisfy, which means that he endorses (unqualified) impossibilism. Pereboom sets his view apart from G. Strawson’s in part by insisting that a being with broadly libertarian, “law-overriding” agent-causal powers could (i.e. in some possible world does) satisfy the source condition on free will and act freely (Pereboom 2001, pp. 85–86, 128). Notably, Pereboom also accepts free-will denialism, the view that *we* (i.e. members of the class *normal humans*) do not have free will, by reasoning which seems to commit Pereboom to anthropocentric impossibilism (for full discussion, see Mickelson “Hard Times for Hard Incompatibilism,” ms.)
- 15 It should now be clear that the impossibilist, when faced with the question “Why can’t a normal person act freely assuming determinism is true?,” cannot adequately answer the challenge by saying “Well, because the actions of such a person would be *completely settled by factors beyond his control*” (e.g. Capes 2019). The challenge posed by E2 is the problem of *pinpointing* the factor(s) which make a person unfree when determinism is true. At best, an allusion to unidentified freedom-undermining “factors” indicates that impossibilism is not a *brute fact* – but this simply confirms the legitimacy of E2 without providing a solution to it.
- 16 This interpretation stands in sharp contrast to the standard narrative within the moral luck literature, according to which the traditional problem of determinism is equivalent to what moral-luck theorists would call the problem of antecedent causal luck, where “causal luck” is understood narrowly (and, I think, mistakenly) as *deterministic* antecedent causal luck (e.g. Latus 2000, p. 167, n. 6). For full discussion, see Mickelson 2019b.
- 17 Notably, moral-luck theorists use the term ‘control’ (e.g. in statements of the “control principle” generates the paradox of moral luck) to pick out the same type of basic-desert-grounding control that most free-will theorists now pick out with the term ‘free will’ (Mickelson 2019b).
- 18 I say “inspired by” because Nagel’s original taxonomy of luck/factors beyond our control was unprincipled: Nagel’s categories were neither mutually exclusive nor collectively exhaustive (he did not suggest they were) in ways that make it unsatisfactory for use in a discussion of rival solutions

to E2 (Mickelson 2019b). Unfortunately, there is no way to capture the overlapping metaphysical problems in the literatures on free will, moral luck, and constitutive luck without making some adjustments to (or completing replacing) the distinctive jargon used in each. For those who resist updating Nagel's terms/categories vis-a-vis moral luck, they are welcome to apply any labels they prefer.

- 19 I understand constitutive luck broadly to include both a lack of control over actual and modal facts about a person. On this understanding, we may distinguish the type of constitutive luck one suffers from when the endowment one has in the actual world is beyond one's control ("endowment luck") and the type one suffers when their essential properties, roughly the properties that underwrite one's cross-world identity, if there are any such properties. That is, it might be that one's entire "modal profile" (roughly one's every possible endowment) is beyond one's control ("modal profile luck"). Following this line of thought, the problem of free will quickly becomes a debate about the metaphysics of personal identity rather than causation or the laws of nature.
- 20 By endorsing James's definition of 'determinism', I have in effect defined 'determinism' to denote the thesis that the naturalistic factors which account for the evolution of the physical universe (laws, causation, or the like) are *strong*, *deterministic*, and *unconditional*; philosophers who think we should add some kind of *ceteris paribus* clause to the traditional (Jamesian-style) statements of determinism are, in effect, using 'determinism' to pick out the slightly different thesis that the relevant evolutionary factors are strong, deterministic, and *conditional* (Mickelson 2019a and Mickelson "Hard Times for Hard Incompatibilism" ms.; see also n. 4 *supra*). Given the latter definition, determinism may be true even in a world at which God regularly intervenes in the natural order – which is interesting because it means that a libertarian might argue that God's intervention introduces freedom-relevant actual-sequence leeway into the world even though "determinism" is true (e.g. Stone 1998). The fact that the latter definition of 'determinism' blurs the line between compossibilists and broadly libertarian accounts of free will speaks against its use.
- 21 Notably, some philosophers have advanced a doctrine paradoxically called "Humean determinism" (Beebe and Mele 2002). I say "paradoxically" because so-called Humean determinism may be true in a world with weak laws of the sort the permit unlimited moment-to-moment actual-sequence leeway in the world (as depicted in van Inwagen's forking-paths diagram earlier). This means that the phrase "Humean determinism" refers to a paradigmatic doctrine of *indeterminism* by any traditional definition of the term 'determinism'. Since this is an odd development, the reader may wonder how such a twist of terminology might have arisen. In brief, the term "Humean determinism" was introduced when it was noted that a certain logical entailment thesis that van Inwagen had referred to by the name 'determinism' (because the entailment thesis served, in logic-text proofs, as a proxy for the metaphysical doctrine depicted in his no-forking-paths diagram) was open to a broadly Humean interpretation. Early commentators took the Humean interpretation of the *proxy thesis* as evidence that there is a Humean interpretation of *traditional zero-leeway determinism*. However, the availability of a Humean interpretation of van Inwagen's *proxy thesis* is better taken as evidence that this proxy failed to capture its target. On the latter line of reasoning, van Inwagen extended the term 'determinism' to his proxy thesis *by mistake*, and it is upon this mistake that the unhappy notion of "Humean determinism" took root. (For a full discussion, see Mickelson "Humean-law Determinism, Humean-law Compatibilism, and The Consequence Argument," ms.)
- 22 Standard abductive reasoning supports the following inference: If (contrary to what libertarian accounts of free will propose) indeterminism doesn't "help" a person to act freely, then (contrary to what incompatibilists claim) determinism doesn't "hurt" – and, indeed, whether determinism is true or false is *irrelevant* to whether we have free will (Mickelson 2019a, 2019b, and "Hard Times for Hard Incompatibilism", ms.).
- 23 Thomas Nagel expresses similar views in his discussions of the paradox of moral luck (Mickelson 2019b).
- 24 Carolina Sartorio comes close to endorsing a strict circumstantial luck solution to E2 when she suggests that determinism would pose no threat to free will but for the fact that we are "causally impotent" vis-à-vis the events in the remote past when determinism is true, but she might

also be interpreted as endorsing a causal/circumstantial hybrid solution (which is distinctive because of her unusual emphasis on the freedom-undermining role played by our lack of control over circumstances in the remote past) (e.g. Sartorio 2016, pp. 151–152). We might classify those impossibilists who contend that past circumstances (apart from such things as the laws of nature and one’s constitutive properties) pose a distinctive threat to free will as *circumstantial luck impossibilists*.

- 25 Philosophers who lack a firm grasp on this in-house debate regarding the solution to E2 often assume that an impossibilist may accept the conclusion of every argument against compatibilism, but this is a mistake. For example, Gregg Caruso seems to endorse both Pereboom’s “hard incompatibilism” and Levy’s “hard luck” impossibilism (e.g. Caruso 2019; Dennett and Caruso 2021, p. 176; see also Mickelson 2021b), but this position (taken at face value) is untenable. While both Pereboom and Levy are impossibilists, they endorse rival solutions to E2 of the explanation problem. Levy – as Caruso notes (Dennett and Caruso 2021, p. 196) – self-identifies as a *compatibilist* to indicate that he rejects the “incompatibilist” thesis that deterministic laws/causation poses a distinct threat to free will (Levy 2011, p. 1, n.1). In more generic language, one cannot accept both (i) Pereboom’s incompatibilism (according to which the truth of determinism is negatively, though *not spuriously*, correlated with the existence of free human agents), and (ii) Levy’s anti-incompatibilist view (which I consider a fleshing out of G. Strawson’s position, e.g. Mickelson 2019b) that determinism is negatively, but only *spuriously*, correlated with the existence of free human agents.
- 26 The notion of sourcehood which drives G. Strawson’s constitutive-luck source impossibilism also fits well with Nagel’s observation that many people are paradoxically committed to the existence and impossibility of moral luck (Mickelson 2019b). It is not at all clear, however, that Pereboom’s broadly libertarian characterization of the source condition fits with his preferred incompatibilist (a.k.a. causal luck impossibilist) solution to E2 (Mickelson “Hard Times for Hard Incompatibilism,” ms.).
- 27 In a recent “glossary for the uninitiated,” Alfred Mele asserts that he is “following standard practice” when he defines ‘incompatibilism’ to pick out a perhaps spurious impossibility claim (e.g. Mele 2019a, p. 1, n.1; see also Mele Ch. 31 this volume); indeed, he asserts that those who use the term in any other way are using it in a “nontraditional” (Mele 2017, p. 6, n. 4) or “nonstandard” way for which he has “never had any use” (Mele 2019b, p. 3, n. 1). Mele’s failure to provide evidence for these claims is notable, for the claims appear to be false. Philosophers regularly use ‘incompatibilism’ (following Lehrer) to pick out a type of explanatory-relevance view, that is, a species of causal luck impossibilism, as opposed to the relatively modest, non-explanatory thesis of mere impossibilism. Given such facts, it is puzzling that Mele has singled out only two philosophers by name – Mickelson (Mele 2017, p. 6, n. 4) and Levy (Mele Ch. 31 this volume) – as examples of philosophers who use the term ‘incompatibilism’ differently than he does, that is, to refer to something other than mere impossibilism. After all, Mele’s comments about ‘incompatibilism’ commit him to the view that Keith Lehrer (who coined the term), Derk Pereboom (2001, 2014), Randy Clarke (2003), Carolina Sartorio (2016), Kadri Vihvelin (2008, 2013), Michael McKenna (2010), John Martin Fischer and Mark Ravizza (1998, p. 151) and many other leading figures in the debate who regularly use the term ‘incompatibilism’ to denote causal luck impossibilism are guilty of using that term in a nonstandard/nontraditional way. Moreover, Mele’s claims about his own use of the term ‘incompatibilism’ are also problematic. Mele used the term ‘incompatibilism’ to pick out a species of causal luck impossibilism when he claimed that Pereboom’s Four-Case Argument “fails as an argument for incompatibilism” (Mele 2005, p. 80; 2006, pp. 144, 189; see also Mele 2008, p. 278). Mele’s attack on Pereboom’s best-explanation reasoning narrowly targets the explanation step of Pereboom’s argument, leaving its counterexample step and generalization step (which supports an impossibilist solution to the correlation problem) untouched (Mickelson 2017). Mele’s treatment of the term ‘incompatibilism’ has also contributed to the common misconception that Mele’s revised Zygote Argument aims to support the same explanatory conclusion as its predecessor, the Four-Case Argument (e.g. Mele 2005, 2006, 2008,

2013, 2017, 2018, 2019b, 2022); in fact, the former concludes to mere impossibilism but the latter concludes to causal luck impossibilism. Notably, Mele has introduced no new terms which would allow him (or his interlocutors) to easily distinguish between the relatively modest correlation-claim conclusion of his *revised* Zygote Argument (Mele 2013, 2017, 2018, 2019b, 2022) and the familiar explanatory conclusion of Pereboom's Four-Case Argument (see Mickelson 2015b, 2017, and 2021a for discussion); since both aim to support *at least* impossibilism, Mele classifies both as "arguments for incompatibilism" and leaves it at that. Finally, because Mele has used different definitions of the term 'incompatibilism' in different dialectical contexts – for example, when he is advancing a criticism of Pereboom's argument (Mele 2005) versus when he is responding to a structurally identical criticism of his Zygote Argument (e.g. Mele 2013; Mickelson 2015b, 2021a) – Mele is vulnerable to the charge of sophistical *motte-and-baileying* on the term 'incompatibilism' and its cognates (Mickelson "Motte-and-Bailey Incompatibilism," ms.).

- 28 For an interesting discussion of ongoing disputes about the demarcations between these relations, see Wirling 2020.
- 29 I have been told (in public discussion and private correspondence) that most of the words/phrases listed here are ambiguous in natural language, such that one may use most of the phrases listed here without indicating that an antagonistic relevance relation holds between things. I doubt this is correct, but I will not challenge the point here. I raise the issue only to make the reader aware of the fact that at least some mainstream analytic philosophers do not use these terms to indicate the presence of a relevance relation.
- 30 In this chapter, I use the term 'proposition' to pick out abstract truth-value-bearing entities. While I recognize the distinction between propositions and statements, I will treat the distinction loosely here because two (or more) statements may stand in any of the logical relations described here, and likewise two (or more) propositions may stand in any of the logical relations described here.
- 31 I purposely avoid language which suggests that every relevance relation is a type of correlation relation (e.g. I avoid phrases such as 'relevance correlation'). I think this way of speaking would exacerbate current confusion by needlessly blurring the sharp distinction between perhaps spurious (i.e. perhaps *non-relevance*) correlation relations and (always, by definition, non-spurious) relevance relations.
- 32 One may draw upon the recent "Dependence Solution" to the problem of free will and foreknowledge to flesh out the asymmetrical dependence relationship of God's foreknowledge on our actions, such that God's knowledge of the future is not a threat to free will (e.g. Hunt and Zagzebski 2022) – even if, as I have stipulated here, something else (determinism-related causal factors beyond one's control) in the world is "doing the work" of destroying free will. (Notably, use of terms 'incompatibility' and 'incompatibilism' has spread to the free will/foreknowledge literature, where the terms are problematically ambiguous in the way that they are in the free will/determinism literature.)
- 33 I do not find it helpful to use the term 'impossibility' to refer to the strict logical inconsistency relation; for those who do, I strongly suggest using the qualifier 'logical'. Failing to do so may lead to confusion, for example, between questions/views about how the *truth-values* of certain propositions are related (if at all) across possible worlds, and questions/views about the possible *co-existence* of a specific proposition (i.e. some truth-value-bearing entity) and some other thing (e.g. another proposition, some other type of abstract object, some physical object, etc.) should be carefully distinguished from questions. For example, to say "Determinism is *logically impossible* (a.k.a. strictly logically inconsistent with) the thesis that someone has free will" is not equivalent to saying "Determinism is *metaphysically impossible* with free will." The latter is a claim about the possible co-existence of the thing picked out by the term 'free will' (which does not have the ontological status of a proposition) and the doctrine of determinism (which has the ontological status of a proposition); this metaphysical impossibility claim says nothing about the *truth-value* of determinism. The former claim, by contrast, assumes the existence of two propositions and asserts something about the truth-values of two propositions. Notably, one may easily reject the latter (metaphysical impossibility) claim without taking a stand on the former (strict logi-

- cal inconsistency/logical impossibility) claim – and, indeed, without taking a stand on any of the central questions in the free-will debate.
- 34 Restated in the language of possible worlds, we may say that statements are logically inconsistent if and only if they cannot be true at the same time *at the actual world* (irrespective of the reason/reasons why). For examples of this generic, non-relevance definition of ‘inconsistency’, see, for example, Bergmann et al. 1990, pp. 2, 16; Tidman and Kahane 2003, p. 16. Given this definition of ‘logical inconsistency,’ both contrary statements (i.e. those which can both be false but cannot both be true) and contradictory statements (i.e. those which cannot both be true and cannot both be false) qualify as logically inconsistent statements – even when the contrariety relation is rooted in the semantic content of the two statements and not in their logical form, for example, “The Taj Mahal is pink all over” and “The Taj Mahal is blue all over” (Layman 1999, p. 144). As such, I consider definitions of ‘logic inconsistency’ on which two propositions/statements are inconsistent if and only if their conjunction entails a contradiction (e.g. Barker 1985, p. 348) to be overly narrow for general purposes, even if there are contexts in which this type of stipulative/technical definition is useful. It is worth noting that introductory logic texts which focus primarily on propositional/sentential logic often describe “truth-table tests” for inconsistency and narrowly define ‘inconsistency’ in terms of such tests (e.g. Lemon 1992, p. 69; Hurley 1994, pp. 322–323, 327–328; Hurley and Watson 2018, pp. 360, 724). The truth-table method is designed to identify a conflict in the respective logical forms of two (or more) statements, where this syntactic conflict alone ensures that the statements cannot be true at the same time (i.e. there is no possible assignment of truth-values to the components of the statements on which the lines under the main operator of each statement has the truth-value true). While such truth-table tests are sometimes said to reveal “truth-functional inconsistency” (Bergmann et al. 1990, pp. 73–75), it seems preferable to say that such tests identify a *syntactic logical incompatibility* between two (or more) statements, for the same reason it is preferable to avoid speaking of a “causal correlation” when one means to say that a given correlation is due to a direct causal relation between the correlates (i.e. we have reason to avoid treating causal relations and relevance relations more generally as special types of correlation relations). For present purposes, I set aside uses of the term ‘inconsistency’ in metatheory (see, e.g., Lemon 1992, pp. 68, 75).
 - 35 Since the material conditional ($p \rightarrow \sim q$) is true so long as its antecedent has the truth-value *false* and/or its consequent has the truth-value *true*, it is also logically equivalent to the disjunction ($\sim p \vee \sim q$).
 - 36 The term ‘logical incompatibility’ is not a technical term in classical logic – indeed, this term rarely mentioned, let alone defined as a technical term, in any standard classical logic text. Presumably this is because the standard tools of classical logic cannot track relevance relations. I take it that recent work on relevance logics (Mares 2020), grounding relations between propositions (e.g. Fine 2012a; 2012b), and the logic of ‘because’ (e.g. Schnieder 2011) is aimed (in part) at fleshing out the sorts of logical relevance relations which fall under (what I am calling) logical incompatibility and strict logical incompatibility.

Bibliography

- Anderson, M.B. (2019). Moral luck as moral lack of control. *Southern Journal of Philosophy* 57 (1): 5–29.
- Ayer, A.J. (1954/2009). Freedom and necessity. In: *Free Will*, 2e (ed. D. Pereboom), Oxford: Oxford University Press. 139–147.
- Bailey, A. (2013). Incompatibilism and the past. *Philosophy and Phenomenological Research* 85 (2): 351–376.
- Baker, L.R. (2006). Moral responsibility without libertarianism. *Noûs* 40 (2): 307–330.
- Barker, S. (1985). *The Elements of Logic*, 4e. New York: McGraw-Hill Book Company.
- Beebe, H. and Mele, A. (2002). Humean compatibilism. *Mind* 111 (442): 201–223.

- Bergmann, M., Moor, J., and Nelson, J. (1990). *The Logic Book*, 2e. New York: McGraw-Hill Publishing Company.
- Bobzien, S. (1998). *Determinism and Freedom in Stoic Philosophy*. Oxford: The Clarendon Press; Oxford University Press.
- Campbell, J. (2007). Free will and the necessity of the past. *Analysis* 67 (2): 105–111.
- Campbell, J. (2008). Reply to Brueckner. *Analysis* 68 (3): 264–269.
- Campbell, J. (2010). Incompatibilism and fatalism: reply to Loss. *Analysis* 70: 71–76.
- Campbell, J. (2011). *Free Will*. Massachusetts: Polity Press.
- Capes, J. (2019). What the consequence argument is an argument for. *Thought: A Journal of Philosophy* 8 (1): 50–56.
- Caruso, G. (2019). A defence of the luck pincer: why luck (still) undermines moral responsibility. *Journal of Information Ethics* 28 (1): 51–72.
- Chisholm, R. (1964/2009). Human freedom and the self? In: *Free Will*, 2e (ed. D. Pereboom), 172–195. Hackett Publishing Company.
- Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.
- Cornman, J. and Lehrer, K. (1968). *Philosophical Problems and Arguments: An Introduction*. New York: Macmillan Company.
- Dennett, D. (2003). *Freedom Evolves*. London: Penguin Books.
- Dennett, D. and Caruso, G. (2021). *Just Deserts: Debating Free Will*. Cambridge: Polity Press.
- Fine, K. (2012a). Guide to ground. In: *Metaphysical Grounding* (ed. F. Correia and B. Schnieder), 37–80. Cambridge: Cambridge University Press.
- Fine, K. (2012b). The pure logic of ground. *Review of Symbolic Logic* 5 (1): 1–25.
- Fischer, J.M. (1994). *The Metaphysics of Free Will: An Essay on Control*. Cambridge: Blackwell Publishers.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility (Cambridge Studies in Philosophy and Law)*. Cambridge: Cambridge University Press.
- Franklin, C.E. (2011). The problem of enhanced control. *Australasian Journal of Philosophy* 89 (4): 687–706.
- Griffith, M. (2017). Major positions in the free will debate. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), 1–3. New York: Routledge.
- Haji, I. and Cuypers, S. (2006). Hard- and soft-line responses to Pereboom's four-case manipulation argument. *Acta Analytica* 21: 19–35.
- Hartman, R.J. (2017). In *Defense of Moral Luck: Why Luck Often Affects Praiseworthiness and Blameworthiness*. New York: Routledge University Press.
- Hermes, C. (2013). Truthmakers and the direct argument. *Philosophical Studies* 167 (2): 401–418.
- Hermes, C. and Campbell, J. (2012). More trouble for direct source incompatibilism: reply to Yang. *Acta Analytica* 27 (3): 335–344.
- Hunt, D. and Zagzebski, L. (2022). Foreknowledge and free will. In: *The Stanford Encyclopedia of Philosophy*, (Summer 2022 Edition) (ed. E.N. Zalta). <https://plato.stanford.edu/archives/sum2022/entries/free-will-foreknowledge/>. (Accessed on 22nd September 2022).
- Hurley, P. (1994). *A Concise Introduction to Logic*, 5e. Belmont, CA: Wadsworth Publishing Company.
- Hurley, P. and Watson, L. (2018). *A Concise Introduction to Logic*, 13e. Belmont, CA: Cengage Learning, Inc.
- James, W. (1884/1897). The Dilemma of determinism. An address to the Harvard divinity students, published in the Unitarian review for September 1884. *Republished in: The Will to Believe and Other Essays* (1897). London: Longmans, Green, and Co.
- Kane, R. (1996). *The Significance of Free Will*. Oxford: Oxford University Press.
- Kane, R. (1999). Responsibility, luck, and chance: reflections on free will and indeterminism. *The Journal of Philosophy* 96 (5): 217–240.
- Kane, R. (2002). Introduction: the contours of the contemporary free will debates. In: *The Oxford Handbook of Free Will* (ed. R. Kane), 1–41. Oxford: Oxford University Press.
- Kane, R. (2004). Agency, responsibility, and indeterminism: reflections on libertarian theories of free will. In: *Freedom and Determinism* (ed. J., Campbell, M. O'Rourke and D. Shier), 70–88. Cambridge: MIT Press.
- Kane, R. (2009). Free will and the dialectic of selfhood: can one make sense of a traditional free will requiring ultimate responsibility? *Ideas y Valores* 58 (141): 25–43.

- Kane, R. (2016). On the role of indeterminism in libertarian free will. *Philosophical Explorations* 19 (1): 2–16.
- Kane, R. (2019). The complex tapestry of free will: striving will, indeterminism and volitional streams. *Synthese* 196 (1): 145–160.
- Latus, A. (2000). Moral and epistemic luck. *Journal of Philosophical Research* 25 (1): 149–172. Available online: <https://www.iep.utm.edu/moralluc> ISSN 2161-0002.
- Latus, A. (2001). Moral luck. In: *The Internet Encyclopedia of Philosophy* (ed. J. Feiser). ISSN 2161-0002, <https://iep.utm.edu/moralluc/>. (Accessed in September 2022).
- Layman, C.S. (1999). *The Power of Logic*. Mountain View, CA: Mayfield.
- Lehrer, K. (1960). Ifs, cans, and causes. Dissertation, Brown University. Providence: ProQuest/UMI. (Publication No. AAT: 6205755.)
- Lemon, E.J. (1992). *Beginning Logic*. Indianapolis: Hackett Publishing.
- Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.
- Mares, E. (2020). Relevance logic. In: *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), (ed. E.N. Zalta), <https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=logics-relevance>. (Accessed in May 2021).
- McKenna, M. (2010). Whose argumentative burden, which incompatibilist arguments?—getting the dialectic right. *Australasian Journal of Philosophy* 88 (3): 429–443.
- McKenna, M. and Pereboom, D. (2016). *Free Will: A Contemporary Introduction*. New York: Routledge.
- McPherson, T. (2021). Supervenience in ethics. In: *The Stanford Encyclopedia of Philosophy*, (Summer 2021 Edition) (ed. E.N. Zalta), <https://plato.stanford.edu/archives/sum2021/entries/supervenience-ethics>. (Accessed in September 2022).
- Mele, A. (2005). A critique of Pereboom’s ‘four-case argument’ for incompatibilism. *Analysis* 65 (1): 75–80.
- Mele, A. (2006). *Free Will and Luck*. Oxford: Oxford University Press.
- Mele, A. (2008). Manipulation and moral responsibility. *Journal of Ethics* 12 (3): 263–286.
- Mele, A. (2012). Another scientific threat to free will? *The Monist* 95 (3) Neuroethics: 422–440.
- Mele, A. (2013). Manipulation, moral responsibility, and bullet biting. *Journal of Ethics* 17 (3): 167–184.
- Mele, A. (2017). *Aspects of Agency: Decisions, Abilities, Explanations, and Free Will*. Oxford: Oxford University Press.
- Mele, A. (2018). Diana and Ernie return: on Carolina Sartorio’s. *Causation and Free Will. Philosophical Studies* 175 (6): 1525–1533.
- Mele, A. (2019a). Free will and moral responsibility: manipulation, luck, and agents’ histories. *Midwest Studies In Philosophy* 43 (1): 75–92.
- Mele, A. (2019b). *Manipulated Agents: A Window to Moral Responsibility*. Oxford University Press.
- Mele, A. (2022). *Free Will: An Opinionated Guide*. Oxford: Oxford University Press.
- Mickelson, K.M. (a.k.a. Kristin Demetriou) (2012). Free Will Fundamentals: Agency, Determinism, and (In)compatibility. Dissertation, University of Colorado, Boulder. https://scholar.colorado.edu/concern/graduate_thesis_or_dissertations/g732d9110.
- Mickelson, K.M. (2015a). A critique of Vihvelin’s three-fold classification. *Canadian Journal of Philosophy* 45 (1): 85–99.
- Mickelson, K.M. (a.k.a. Kristin Demetriou) (2015b). The zygote argument is invalid—now what? *Philosophical Studies* 172 (11): 2911–2929.
- Mickelson, K.M. (a.k.a. Kristin Demetriou) (2017). The manipulation argument. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), New York: Routledge.
- Mickelson, K.M. (a.k.a. Kristin Demetriou) (2019a). Free will, self-creation, and the paradox of moral luck. *Midwest Studies in Philosophy* 43 (1): 224–256.
- Mickelson, K.M. (a.k.a. Kristin Demetriou) (2019b). The problem of free will and determinism: an abductive approach. *Social Philosophy & Policy* 36 (1): 154–172.
- Mickelson, K.M. (a.k.a. Kristin Demetriou) (2021a). The zygote argument is still invalid: so what? *Philosophia* 49 (2): 705–722.

- Mickelson, K.M. (a.k.a. Kristin Demetriou) (2021b). Just deserts: debating free will. *International Journal of Philosophical Studies* 29 (3): 408–412.
- Nagel, T. (1976). Moral Luck II. *Proceedings of the Aristotelian Society, Supplementary Volumes* 50: 137–151.
- Nagel, T. (1979). *Mortal Questions*. New York: Cambridge University Press.
- Nagel, T. (1986). *The View from Nowhere*. Oxford: Oxford University Press.
- Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research* 73 (1): 28–53.
- Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2005). Defending hard incompatibilism. *Midwest Studies in Philosophy* 29 (1): 228–247.
- Pereboom, D. (2009). The stoics. In: *Free Will*, 2e (ed. D. Pereboom), 5–16. Indianapolis: Hackett Publishing Company, Inc. Originally published in *Hellenistic Philosophy* (ed. and trans. B. Inwood and L.P. Gerson). Indianapolis: Hackett Publishing Company (1988).
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
- Perry, J. (2004). Compatibilist options. In: *Freedom and Determinism* (ed. J. Campbell, M. O'Rourke and D. Shier), 231–254. Cambridge: MIT Press.
- Sartorio, C. (2016). *Causation and Free Will*. Oxford: Oxford University Press.
- Schnieder, B. (2011). A logic for 'because'. *The Review of Symbolic Logic* 4 (3): 445–465.
- Sehon, S. (2010). A flawed conception of determinism in the consequence argument. *Analysis* 71 (1): 30–38.
- Shabo, S. (2011). What must a proof of incompatibilism prove? *Philosophical Studies* 154 (3): 361–371.
- Statman, D. (2019). The definition of luck and the problem of moral luck. In: *The Routledge Handbook of the Philosophy and Psychology of Luck* (ed. I.M. Church and R.J. Hartman), 195–205. New York: Routledge.
- Stone, J. (1998). Free will as a gift from god: a new compatibilism. *Philosophical Studies* 92 (3): 257–281.
- Strawson, G. (1986). *Freedom and Belief*. Oxford: Clarendon Press.
- Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies* 75 (1–2): 5–24.
- Strawson, G. (2002). The bounds of freedom. In: *The Oxford Handbook of Free Will* (ed. R. Kane), 441–460. Oxford: Oxford University Press.
- Strawson, G. (2008). The impossibility of ultimate moral responsibility. In: *Free Will*, 2e (ed. D. Pereboom), 289–306. Oxford: Oxford University Press.
- Strawson, G. (2011). Free Will. In: *Routledge Encyclopedia of Philosophy* (ed. E. Craig) London: Routledge. doi: 10.4324/9780415249126-V014-2. <https://www.rep.routledge.com/articles/thematic/free-will/v-2> (Accessed in January 2022).
- Tidman, P. and Kahane, H. (2003). *Logic and Philosophy*, 9e. Belmont, CA: Wadsworth/Thomson Learning.
- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Oxford University Press.
- van Inwagen, P. (1990). Logic and the free will problem. *Social Theory and Practice* 16 (3): 277–290.
- van Inwagen, P. (1999). Moral responsibility, determinism, and the ability to do otherwise. *Journal of Ethics* 3 (4): 343–351.
- van Inwagen, P. (2000). Free will remains a mystery: the eighth philosophical perspectives lecture. *Philosophical Perspectives* 14: 1–19.
- van Inwagen, P. (2017). The problem of $fr^{**}w^{*}ll$. In: *Thinking about Free Will* (ed. P. van Inwagen), 192–209. Cambridge: Cambridge University Press.
- Vihvelin, K. (2008). Compatibilism, incompatibilism, and impossibilism. In: *Contemporary Debates in Metaphysics* (ed. T. Sider, J. Hawthorne and D.W. Zimmerman), 303–318. Malden, MA: Blackwell Publishing.
- Vihvelin, K. (2013). *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. Oxford: Oxford University Press.
- Vihvelin, K. (2022). Arguments for incompatibilism. In: *The Stanford Encyclopedia of Philosophy*, (Fall 2022 Edition) (ed. E.N. Zalta and U. Nodelman), <https://plato.stanford.edu/archives/fall2022/entries/incompatibilism-arguments/>. (Accessed in September 2022).
- Wirling, Y.S. (2020). Is backing grounding? *Ratio* 33 (3): 129–137.

Agent Causation

LEIGH C. VICENS

According to incompatibilists, the sort of free will required for moral responsibility is incompatible with causal determinism. As some see it, if a person's actions are determined by events of the distant past, over which she has had no control, then her actions themselves are not under her control in such a way that she could be responsible for what she does. One might guess that such thinking would lead incompatibilists to conclude that free actions (or certain events such as the formation of intentions that immediately precede them) must be either uncaused, or caused only non-deterministically by prior events. And, indeed, some libertarians (those incompatibilists who believe that humans are sometimes free) develop theories of free will that are non-causal or non-deterministic ("event-causal") in character.

Other incompatibilists, however, insist that neither of these alternatives allows for free will. The problem with non-causal theories, they say, is that *control* is inherently a causal notion; so, if a person's actions are not caused by anything, including the person, then her actions cannot be under her control (O'Connor 2000, p. 25; cf. Clarke 2003, pp. 40, 43; Pereboom 2014, pp. 19–21). And the problem with event-causal libertarian theories is that while they do not provide any *less* control than deterministic theories do, they also don't provide any *more* control. As Derk Pereboom explains it, an action deterministically caused by events beyond an agent's control is not free, and neither is an action not caused by anything at all; but then, an action non-deterministically caused by events beyond the agent's control "exhibits only a combination of the first two types of responsibility-undermining factors," and so cannot be free either (2001, p. 48). Randolph Clarke puts the point somewhat differently: an event-causal libertarian theory "secures the openness ... required for directly free action. But this openness is, figuratively speaking, merely the space in which the agent exercises her active control. It does not augment the active control she exercises but merely allows that there is more than one way in which this active control might be exercised" (2003, p. 96). Likewise, Timothy O'Connor writes, "the free control [event-causal libertarian theories] posit is secured by an *absence*, a removal of a condition (causal determination) ... If there is no means by which I can take advantage of this looser connectivity in the flow of events, its presence can't confer a greater kind of control, one that *inter alia* grounds moral responsibility for the action and its consequences" (2008, pp. 192–193). While

indeterminism “provides an opening in which the agent might freely select one option from a plurality of real alternatives, it fails to introduce a causal capacity that fills it” (O’Connor 2011, p. 311).

As this last quotation suggests, the incompatibilists under consideration here maintain that in order to secure the kind of control essential to free will, a novel causal capacity must be introduced. That capacity, called “agent-causation,” is one which gives the agent *her-self* the power to cause her actions. Causation by an agent, on this view, is not reducible to causation by any events, including those “involving” the agent, such as the occurrence of some consideration in the agent’s mind in favor of acting in a particular way. Rather, agent-causation is strictly and literally causation by a substance – in this case, the person who is acting. As the positing of agent-causal power is motivated by the perceived failure of non-causal and event-causal libertarian theories to secure the control necessary for freedom – what I shall call, following Franklin (2011), “the problem of enhanced control” – its exercise is supposed to be “intrinsically a direct exercise of control over one’s behavior” (O’Connor 2000, p. 61). Some theorists argue that agent-causal power could be exercised in a deterministic world, and develop agent-causal accounts of free will that are compatible with determinism (cf. Ned Markosian 1999; Nelkin 2011). Nevertheless, the majority of agent-causal theorists are incompatibilists, maintaining (as suggested in the quotations above) that indeterminism provides the “opening” or “space” necessary for persons to exercise the control that agent-causal power makes possible. Thus, in what follows I will focus on agent-causal *libertarian* theories.

Agent-causal libertarian theories come in different varieties, depending on how they understand the relationship between agent- and event-causation. Their differences make them subject to some distinct objections. Below I review the two main contender theories and summarize their different liabilities, before going on to discuss further objections that have been raised more generally to the conceivability and empirical plausibility of agent-causation.

According to the more traditional agent-causal account recently developed by Timothy O’Connor, a free action is an agent’s causing a certain event: the formation of an intention to act in a particular way. The action that is directly free is thus a *mental* action – a choice or decision – which in normal cases initiates an “extended sequence of event-causal processes” resulting in the agent’s carrying out the intention in further action (2011, p. 315). Crucially, on O’Connor’s account the event that an agent freely causes is *not* caused by any previous events, including those involving the agent, such as the agent’s consideration of a reason for action. Still, on O’Connor’s view, reasons figure into the explanation of the agent’s free action, since the event that an agent causes is an intention to act *for* a particular reason, such as to satisfy some desire (2001, p. 86). While O’Connor insists that an agent’s reasons cannot be *causes* of her free action, on his account the reason for which the agent intentionally acts, as well as other reasons which may not figure into her intention, still affects the evolving probability that the agent will form the particular intention that she does, by giving her “varying propensities towards different outcomes” (2005, p. 216). When reasons have such an effect on the probability of the action that the agent actually performs, the agent may be said to act *on* those reasons.

O’Connor’s attempt to explicate the role of reasons in free action, to show that agent-caused behavior can be rationally explicable, raises a number of questions. One may wonder, first, how it can be that an agent’s reasons for various possible actions can affect the probability that she performs those actions, if the occurrence or consideration of those reasons, as events, cannot *cause* her free action. What is the non-causal way in which reasons have such an effect? O’Connor merely says that events which influence the probabilities of various possible actions “[affect] the propensities of the agent-causal

capacity itself" (2008, p. 198). But it is unclear how they could have this effect without being at least partial causes of the action which results from the exercise of this agent-causal capacity.

Alfred Mele and Randolph Clarke raise further objections to O'Connor's account, focusing on his claim that what makes it the case that an agent acts for a reason is that the reason figures into the content of the agent's intention. Mele notes, in responding to a similar claim made by R. Jay Wallace, that figuring into the content of the agent's intention is not *necessary* for a reason to count as one for which the agent decided. He writes:

We consider many reasons for and against accepting certain job offers, for example, and sometimes we reach the decisions we do in these cases – and accept or reject a job offer – *for* a whole raft of reasons. It is unlikely that large rafts of reasons can be read off from the contents of our decisions in such cases. It would take a very special mind to represent each member of a large collection of reasons in the content of a decision. If, in cases of this kind, people should say... that what the agent decided was to accept job offer X – and not that what he decided was, for example, to accept X “as a way of” bringing it about that he and his family live in a more attractive part of the world, enabling his children to attend better schools inexpensively, improving his family's job prospects, reducing his teaching load, increasing his salary, and so on – special grounds need to be offered for holding that, even in relatively simple cases, (partial) representations or reflections of each of the reasons for which agents decide and act as they do uniformly enter into the content of decisions. (2003, p. 43)

Clarke further notes that figuring into the content of the agent's intention is not *sufficient* for a reason to count as one for which the agent decided, since the content of a decision “can encompass a goal that, despite what the agent might believe, is not in fact a reason for which that decision is made” (2011, p. 343).

Of course, in order for these objections to be successful, Mele and Clarke must show that a decision might count as a *free* one, even though the agent's intention does not encompass all, or some, of the reasons for which she is actually deciding. O'Connor might take issue with Clarke's objection, since O'Connor thinks that at least some “knowledge of the factors motivating one to act” is necessary for free will (2005, p. 215). Yet Mele's point, that one might act (freely) for a number of reasons without all of them figuring into the content of one's intention, still seems to have some force against O'Connor's view.

Considerations such as these may lead one to conclude that a better account of the role of reasons in free actions is simply a *causal* one. Such is the approach Clarke has taken in his “integrated” agent-causal account. On this view, every free action is caused *both* by an agent, *and* by certain agent-involving events. Of course, if an agent and her reasons caused the agent's action independently of each other, then the action would be over-determined, so that one of the causes would be unnecessary for the action to occur. But Clarke theorizes that there must be a law of nature tying agent-causes and event-causes together, such that neither can cause the resulting action without the other (2003, p. 145).

Clarke maintains that his view secures “both the (literal) origination that is provided by traditional agent-causal accounts and what is needed for acting for reasons and reasons-explanation” (2003, p. 136). Moreover, he takes it as a virtue of the integrated agent-causal account that according to it, free human action is “less fundamentally different from phenomena that appear elsewhere in nature” since it is “thoroughly integrated into the nexus of event causes.” The account can thus make sense of how a person's free choices are influenced by her genetic makeup, past experiences, and current environment, and so allow for scientific investigation and explanation of human behavior (2003, p. 137).

Clarke's integrated account seems to avoid the difficulties described above facing the more traditional agent-causal account, which denies that events may cause free choices; but it has been subject to objections of its own from defenders of those traditional accounts. One objection raised by O'Connor and John Churchill begins by noting that for agent-causation to serve its purpose as a solution to the problem of enhanced control, an agent's causing of her own action must have some kind of explanatory priority over any event's causing it. For suppose, to the contrary, that the agent did not have any such priority, but that instead the event-cause was explanatorily prior; then the account wouldn't be an improvement on a kind of deterministic agent-causal theory, according to which the agent causally contributes to some effect but, rather than being the ultimate determiner of her action, is dependent on events beyond her control to determine which action she performs. But in fact, O'Connor and Churchill argue, on Clarke's view the agent-cause does not have explanatory priority, since "we cannot say that an instance of event causation occurred *because* an instance of agent causation occurred in any strong sense than we can say that an instance of agent causation occurred because an instance of event causation occurred" (2004, p. 245). Indeed, they contend, on Clarke's view the event-cause seems to have priority, since it is possible for events to cause a (non-free) action without the agent making a causal contribution, while it is *not* possible for an agent to cause an action (whether free or not) without the causal contribution of some event (2004, pp. 246–247).

In response to this second problem raised by O'Connor and Churchill – that on Clarke's account, the event-cause seems to have explanatory priority over the agent-cause – Clarke might respond that, while it is possible for an event to cause an action without the agent making a causal contribution, this is true on most versions of the traditional account as well (including O'Connor's), and simply serves to highlight the *difference* between non-free and free action. That is to say, non-free action is not free precisely *because* it is caused by some events without the agent making a causal contribution, and so the agent lacks sufficient control over it.

In response to the first problem raised by O'Connor and Churchill, Clarke maintains that his account does in fact secure an "asymmetry in favor of the agent-cause":

Given that an agent has what it takes to act freely, a decision cannot be made unless the agent causes it. But which decision is made need not similarly depend on which psychological causes are effective. Allen might have several distinct reasons for lying and – in addition to its being open to him to decide to tell the truth – it might be open to him to decide to lie for one of these reasons and open to him to, instead, decide to lie for another of these reasons. Which decision he makes, then, might depend on him in a way in which it does not depend on which psychological events cause it. (2011, p. 344)

Clarke is right that there is an asymmetry between agent- and event-causes on his view, since the agent *must* be a cause of any free action she performs, while for some free action she might perform, it may be that no *particular* reason must be the cause (since there may be more than one reason that could cause it). Still, it is unclear whether this asymmetry grounds an explanatory priority of the sort O'Connor and Churchill demand. In the passage cited above, Clarke makes it seem that, in the case of an action which has two possible event-causes, it's *up to the agent* which event actually causes the action. But it would not seem to follow, from the fact that there is more than one possible event-cause, that it "depends" on the agent in such a responsibility-conferring way which event actually does cause the action.

A related objection concerns how a view like Clarke's, which aims to thoroughly integrate agent-causation "into the nexus of event causes," can accommodate the idea that event-causes are *law-governed*. If there are universal laws of nature of a statistical nature which fix the probability at every time of every possible event that might occur at every later time, how can agents, *qua* substances, have a kind of power distinct from these events to affect how things will turn out? Derk Pereboom lays out the problem here as a kind of dilemma: either agent-causal power is reductive, or it is not. What it means for agent-causal power to be *reductive* is for it to result "from a wholly microphysical constitution by virtue of the organization of its constituents" in such a way that "the microphysical level remains wholly governed by the physical laws" (2001, p. 70). If agent-causal power is not reductive in this way – if the microphysical level is not wholly governed by the ordinary laws of physics – then the causal power is *emergent*. Now Pereboom argues that if agent-causal power is reductive, it does not solve the problem that it was proposed to solve – that of enhanced control – and so is no better off than event-causal libertarianism. For every decision an agent makes, and every action she performs, is constituted by events that are either deterministic, or random, or "partially random" – "those for which factors beyond the agent's control contribute to their production but do not determine them, *while there is nothing that supplements the contribution of these factors to produce the events*" (2001, p. 71, emphasis added). In other words, the reductive view allows no role for the agent, *qua* substance, to causally contribute to her actions in a way that allows them to be up to *her*, as their ultimate source.

So, then, the agent-causal libertarian must be an emergentist. But Pereboom reasons, if agents have the ability to act in ways that deviate from the statistical laws governing their microphysical constituents, we should expect to find evidence of such deviations in the form of statistical abnormalities. Otherwise, it would be too wild a coincidence if in the long run the frequencies of choices freely caused by agents turned out to be the same as we would expect to find were there no agent-causes. Yet, Pereboom contends, there is no evidence of such deviation (2001, pp. 82–85).

Clarke has responded to this argument, maintaining that the conformity of agent-caused events to the laws of physics is not a mere coincidence, since "agent causation is nomologically tied to causation by events of certain types," so that if the laws governing those events "derive ultimately from microlevel laws, then the exercise of agent-causal power will have to accord with these microlevel laws" (2003, p. 181, n. 31). Pereboom finds this response unsatisfying, however, charging that "This move would appear to transfer the coincidence to ... a law, which, because it is brute, does not explain the coincidence" (2014, p. 68). Perhaps, though, Clarke means not to "transfer the coincidence" to a law, but to maintain that the statistical laws under consideration do not assign probabilities to the occurrence of events simply on the basis of their potential event-causes, but also on the basis of any potential agent-cause, so that the causal contribution of agents is *already* taken into account by the law. This interpretation is supported by comments Clarke has made in a more recent article:

In contexts in which agents have ... a structured [agent-]causal power, some of the laws concerning which events are likely to follow which others will at the same time set probabilities of agents' causing these outcomes. *For some outcomes, the probabilities of their occurrence will be probabilities of their being agent-caused.* Given agent causation so structured, outcome distributions for agent-caused events would likely conform to probabilities set by the relevant diachronic laws. (2010, p. 397, emphasis added)

This possibility presents a further problem, however. For the laws governing the microphysical constituents of agents – say, the subatomic particles – are supposed to be relatively simple laws that have governed such particles in times long before there were any agents, and in places where no agents exist. It is thus hard to believe that these laws could give distribution outcomes for the exercise of such a complex causal power as that which confers upon an agent the ability to exercise it at will and freely act in a variety of different ways. But if the laws of microphysics don't take into account the probabilities of events being agent-caused, then Pereboom seems right, that it would be a wild coincidence if the frequencies of agents' free actions ended up conforming to those predicted by the laws.

Still, one might question Pereboom's claim that there is no evidence of deviations from the laws governing microphysical events – or, rather, his assumption that *were there such deviations, we should expect to have evidence of them*. Given that the laws under discussion here are assumed to be of a statistical nature, the only way to determine if agent-caused events deviate from them would be to put an agent in exactly the same circumstances many times over and see if the frequencies of her resulting actions conform to the frequencies we would expect on the basis of the laws alone. But this is an experiment we could never carry out, since it would be impossible to guarantee *exactly* the same circumstances each time, including both the microphysical environment and the agent's own states of mind. Clarke takes this fact – that “We never ourselves repeat our decision making in exactly the same states and circumstances” – to mean that the “likelihood of what the distribution of nonaction repetitions would be” should not be thought to “preclude the freedom of choice” (2010, p. 391). But my point here is different: it is that, given we never repeat our decision-making in exactly the same states and circumstances, we cannot know whether our decision-making is anomalous or not. Whether the requirement that agent causes be emergent counts against the plausibility of agent-causal libertarianism thus depends on whether there is some independent reason (besides the lack of evidence for deviations from microphysical laws in the sphere of human behavior) for thinking that these laws are universal, in the sense that they provide distribution outcomes for *all* events.¹

In any case, if agent-causal theorists reject reductionism and embrace emergentism in order to allow for agents the kind of control over their actions essential to freedom, they will again face questions of how certain agent-involving events, such as the agent's consideration of reasons for action, can affect the outcome of the agent's deliberation process. Perhaps the answer is to say (contra O'Connor) that events may be co-causes of an agent's free actions but (contra Clarke) that actions which are co-caused by agents and events are not subject to the probability distributions predicted by the laws governing other events.²

So far I have discussed objections to the two main types of agent-causal libertarian theories, focusing on the relationship between agent-causation and event-causation, including the way in which an agent's states of mind, such as her considerations of reasons for action, may influence her behavior, and the question of whether these views are compatible with a view of microphysical laws as universal – governing even human behavior – though statistical in nature. Two final objections will be considered to the plausibility of agent-causation, one regarding whether there are really two distinct kinds of causation – by agents, and by events – and the other having to do with whether agent-causation, even if real, would really secure more control than event-causation could.

The first objection is that is incredible to suppose that causal power might be exerted in the two different ways that most agent-causal theorists claim: by an object's having a property at a particular time, in the case of event-causation, and by a substance, in the case of agent-causation.

Clarke admits that since causation would not seem to be such a “radically disunified phenomenon,” this objection “counts against the possibility of agent causation” (2003, pp. 208–209). One response to this concern about the uniformity of causal power is to suppose that, in fact, all causation is fundamentally substance-causation, as E. J. Lowe (2008) has argued. Of course, if all causation is substance-causation, then it is hard to see how substance-causation could be what resolves the problem of enhanced control and so enables humans to act freely, since other substances lack this type of control. But perhaps it is something about *agents*, in particular, and not *substances*, in general, that gives humans the power to control their actions. (Lowe suggests that it is our possession of rational powers that gives us the ability to make free choices.) But even setting aside the question of whether those who maintain that all causation is substance-causation could account for the difference between free agents and non-free substances, Clarke argues that if there is ultimately only one kind of causation, it makes more sense to say that it is *event*-causation. For, he reasons, “events, unlike substances, are directly in time, and this allows for a more straightforward accounting for the fact that causes cause their effects to occur at specific times. Further, because events (and not substances) are directly in time, they seem capable of influencing the chances of their future effects in the way that causes should be able to” (2003, p. 209).

O'Connor, who agrees with Clarke that event-causation is real and irreducible to substance causation, finds the “uniformity of causal powers” objection untroubling, however. While acknowledging that there must be unity at some level of abstraction, he asks, “at *what* level of abstraction should the thesis be applied?”

Consider that, in the advent of statistical laws in fundamental physics, many metaphysicians are now comfortable with the notion that there are nondeterministic dispositions varying in strength along a continuum, with deterministic potentialities merely being a limiting case. Consider further that, while properties typically work in tandem toward effects, a natural way of interpreting the phenomenon of radioactive particle decay is as an entirely self-contained process whose timing is radically undetermined by any sort of stimulus event. Finally, some adhere to the truth of ... a view that all or many conscious mental properties are intrinsically intentional while this is true of no physical properties. None of these claims concern free will, and yet all posit a kind of variability in the nature of dispositional properties that warrants classifying them into different basic types. Given these examples, it is hard to see why there may not be a further partition of types of the sort envisioned by the agent causationist. Doubtless there is a unity across these divisions at *some* level of abstraction. But assuming the agent causationist's position is otherwise motivated, he may reasonably contend that it must be sufficiently abstract as to encompass the division his theory requires. Indeed, why may not the unity of *basic* dispositional properties simply consist in their making a net addition to the pool of causal powers? (2011, pp. 319–320)

O'Connor suggests that if the agent-causal theorist's view were *not* otherwise motivated, the concern for uniformity would be some reason to reject the view. This brings us to a final objection to the agent-causal theorist's view: that it makes significant metaphysical and empirical commitments – some quite controversial, as we have seen – without sufficient motivation. In particular, critics maintain that the view does not solve the central problem it was developed to solve: that of enhanced control. Mele sees this problem as a species of what he calls “the problem of present luck.” He notes that O'Connor, among others, describes the failure of event-causal libertarianism in terms of chanciness: on those views, O'Connor says, “There are objective probabilities corresponding to each of the [possible choices], but within those fixed parameters, which choice occurs on a given occasion seems, as far as the agent's direct control goes, a matter of chance” (O'Connor 2001, p. xiii; cited in Mele 2006, p. 53).

Elsewhere O'Connor writes, "Given the causal indeterminist view, if I am faced with a choice between selfish and generous courses of action, each of which has some significant chance of being chosen, it would seem to be a matter of luck, good or bad, whichever way I choose, since I have no means directly to settle which of the indeterministic propensities gets manifested" (2008, pp. 192–193). Likewise, Pereboom describes his own "disappearing agent" objection to event-causal libertarianism as a version of a luck objection (2014, p. 32).³

Mele notes the similarity between O'Connor's way of phrasing the problem and his own "problem of present luck" (2006, p. 66), which he puts in terms of possible worlds: suppose in the actual world a man named "Tim" decides at time *t* to stop working. In another world with the same laws of nature and the same past, Tim decides at *t* to continue working. On the event-causal libertarian picture, there is no difference between the worlds that can account for their divergence; so, the difference between the worlds seems to be just a matter of luck (2006, p. 54; cf. 9). Mele then explains how he thinks the agent-causal view is supposed to avoid the consequence that the difference between the man's decision in the two worlds is a matter of luck, and raises his own objection to this account:

Assume that Tim chose freely in the scenario under consideration. Then, on O'Connor's view, Tim "had the power to choose to continue working or the power to choose to stop, where this power is a power to cause either of these mental occurrences. That capacity was exercised at *t* in a particular way (in choosing to continue working), allowing us to say truthfully that Tim at time *t* causally determined his own choice to continue working"... Suppose that the position reported in the preceding two sentences is true. Why should we suppose that the following cross-world difference is not a matter of chance or luck: that Tim exercised the capacity at issue at *t* in choosing to continue working rather than in choosing to do something else...? (2006, p. 54)

Mele thinks the question is a fair one, since "O'Connor does not place cross-world differences in agents' doings out of bounds in the context of free will; in fact, such differences are *featured* in his objection from chance to event-causal libertarians" (2006, p. 55).

In fact, however, O'Connor does not seem to think that it is the cross-world difference itself that raises a problem for the event-causal libertarian view. In response to Mele, he writes that the problematic sort of luck that he sees involved in event-causal libertarianism "attaches not to a contrastive fact about what occurs, but to the occurrence itself" (2011, p. 325). As I interpret O'Connor here, he is saying that there is a significant difference between the event- and agent-causal libertarians' views of the *actual occurrence* of a free decision, since on the former (but not the latter), an agent is lacking a form of control "that would be exercised were the agent in a maximally direct manner to bring about one or another option – to settle which of her probabilistic dispositions will be manifested on that occasion" (2011, p. 325; cf. Griffith 2005, p. 266). It is this difference – and not the fact that (on both views) nothing *prior* to the time at which the agent makes a decision explains why she decides one way rather than another – which precludes free will in the event-causal libertarian scenario.⁴

Mele finds O'Connor's response unsatisfying, however. As he puts the point in his original criticism of O'Connor's reasoning, it is unclear why we should believe that the kind of power attributed to agents on the agent-causal libertarian view – the power to directly causally determine one's own choices – is sufficient for the power to *freely* make a choice (2006, p. 56; cf., 2014, p. 552). A further response that agent-causal libertarians might make is to say that the agent-causal power they posit just *is* the power necessary for free will and moral responsibility to control one's choices. Indeed, O'Connor suggests as much when defining

agent-causal power in the first place; after describing the problem of luck facing event-causal libertarianism, he writes:

Thus, freedom requires more than indeterministic propensities, more than chance-like outcomes of general tendencies present within the person. Free agents actively determine which tendency will come to fruition on a particular occasion. *We have seen that the only plausible strategy for reducing this capacity to something more basic fails. Thus, some philosophers, myself included, conclude that it is ontologically basic.* (2005, p. 216, emphasis added)

Elsewhere, O'Connor writes, "exerting active power is *intrinsically* a direct exercise of control over one's own behavior. Consequently, the agency theory needn't tell a further story that explains how the agent controls this event itself" (2001, p. 61, emphasis added).

If agent-causal power is just whatever power gives an agent sufficient control over her action to count as free and responsible for it, then of course nothing more can or needs be said in defense of the claim that agent-causal power solves the problem of enhanced control. But then the agent-causal libertarian's "solution" to the problem would seem to be an empty one. As Gary Watson charges, "at best, the term 'agent cause' seems to be just a label for what is needed" (2003, p. 10). To show that this is not an empty solution, the agent-causal theorist must give some account of what the power amounts to, and why this is an improvement over the event-causal picture. In fact, though, the agent-causal theorists already considered *do* give such accounts, laying out just what it means for an *agent*, rather than an *event*, to cause an action. The challenge, in response to the luck objection, is to further explain why this difference makes a difference in the realm of freedom and responsibility. O'Connor, for instance, may need to say to those who do not share his intuition something more than, "The agent's control is exercised not through the efficacy of prior states of action ... but in the action itself ... And because, *ex hypothesi*, it is literally the agent herself generating the outcome, *it is hard to see how the posited form of control could possibly be improved on*" (2011, p. 324, emphasis added). While it may be hard to see how this form of control could be improved on, event-causal libertarians may demand further explanation of why this "improvement" to their view is needed.

In this chapter I have focused on one main argument in favor of the agent-causal libertarian view – that it alone can solve the problem of enhanced control – as well as several objections to the view (and the argument in its favor). A distinct motivation for positing agent-causal power has to do with the phenomenology of agency – that is, the way it *feels*, from "the inside," to act freely. O'Connor, in particular, sees the phenomenology of agency as central to the case to be made for agent-causal libertarianism:

It is not, after all, simply to provide a theoretical underpinning for our belief in moral responsibility that the agency theory is invoked. First and foremost ... the agency theory is appealing because it captures the way we experience our own activity. It does not seem to me (at least ordinarily) that I am caused to act by the reasons which favor my doing so; it seems to be the case, rather, that *I* produce my decision *in view of* those reasons, and could have, in an unconditional sense, decided differently. (1995, find page).

O'Connor cites a number of other philosophers who endorse this picture of the phenomenology of agency, and reasons, "If these largely similar accounts of the experience of action are, as I believe, essentially on target, then it is natural for the agency theorist to maintain that they involve the *perception* of the agent-causal relation" (1995, find page). Other philosophers, however, report a decidedly different phenomenology of agency that

would seem to support an event-causal theory over an agent-causal one (cf. Dennett 1984, p. 78), and still others question the reliability of introspective reports given by philosophers who may introspect “through the lens of their theoretical commitments” (Nahmias et al. 2004, p. 163). Eddy Nahmias and his colleagues have conducted talk-aloud studies on the “folk” – those not influenced by such theoretical commitments – and found “no evidence of agents experiencing themselves as the causal source of their choices” (2004, p. 178), while Terry Horgan et al. maintain that it is possible to affirm the “self as source” phenomenology reported by agent-causal theorists such as O’Connor and yet deny that agent-causation is metaphysically required for this phenomenology to be veridical (2003). While considerations of space preclude a full discussion of the debate about the phenomenology of agency as motivation for agent-causal theories, suffice it to say that further research on this question may be in order.

Notes

- 1 For further discussion of the evidence (or lack thereof) for the claim that physics has established that the statistical laws of physics are universal, such that every event has an objective prior probability of occurring, see Dupré (1996), Loewer (1996), and Buchak (2013). Vicens (2016) also discusses further introspective and sociological evidence (or lack thereof) for thinking that all actions have objective prior probabilities of occurring.
- 2 It should be noted that O’Connor embraces an emergentism of sorts, while still maintaining that free actions have particular objective probabilities of occurring. On his view, agent-causal power is emergent in the sense that “it exerts a causal influence on the micro-level pattern of events that is not reducible to the *immediate* causal potentialities of the subvening properties. The idea is that microphysical properties typically manifest individual potentialities locally – ones that are discernible in fairly isolated contexts – while also carrying a capacity to generate (in tandem with other properties jointly embedded in a properly organized context) emergent features of the system as a whole” (2001, p. 112). Given such emergence, the laws that normally govern microphysical processes may fail “fully to govern these complex systems” (2001, p. 113). Elsewhere O’Connor describes the way an emergent power may exert its influence at the microphysical level as “not a violation of the laws of particle physics but ... a supplement to them” (2008, p. 194). It is unclear to me what it could mean to *supplement* a law that says that particular kinds of events have particular probabilities of occurring in particular circumstances, but in any case Pereboom’s argument against the conformity of (agent-caused) free actions to such statistical laws would seem to have force against O’Connor’s position as well as Clarke’s.
- 3 Christopher Franklin distinguishes what he calls “the problem of enhanced control” from luck objections, since he says the latter are meant to show that indeterminism is incompatible with free will and moral responsibility because it diminishes control, whereas the former “concedes ... that indeterminism does not diminish control ... but questions whether indeterminism can enhance control” (2011, p. 688). Clarke seems to interpret one “argument from luck” along similar lines as Franklin (2003, pp. 77–82). However, since I think O’Connor and Pereboom, at least, understand the main problem with event-causal libertarianism to be a problem with luck – and a problem that agent-causal libertarianism can solve – I will not follow Franklin’s classification here.
- 4 It is sometimes said that the problem with event-causal libertarianism is that nothing prior to the occurrence of the decision “determines” or “settles” whether the decision will occur, and so, that nothing does (cf. Pereboom 2014, p. 32). But event-causal libertarians may respond that, on their view, something *does* settle what will occur – *namely, the event that causes the decision* – just as the agent-causal libertarian will say that it is the *agent* that settles this. If there is a problem with event-causal libertarianism, then, it must be in the proposal that it is an event and not the agent herself that does the settling.

Bibliography

- Buchak, L. (2013). Free acts and chance: why the rollback argument fails. *The Philosophical Quarterly* 63 (250): 20–28.
- Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.
- Clarke, R. (2010). Are we free to obey the laws? *American Philosophical Quarterly* 47 (4): 389–401.
- Clarke, R. (2011). Alternatives for libertarians. In: *The Oxford Handbook of Free Will*, 2e (ed. R. Kane), 329–348. Oxford: Oxford University Press.
- Dennett, D. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press.
- Dupré, J. (1996). The solution to the problem of the freedom of the will. *Nous* 30, *Supplement Philosophical Perspectives* 10: 385–402.
- Franklin, C.E. (2011). The problem of enhanced control. *Australasian Journal of Philosophy* 89 (4): 687–706.
- Griffith, M. (2005). Does free will remain a mystery? *Philosophical Studies* 124 (3): 261–269.
- Horgan, T., Tienson, J., and Graham, G. (2003). The phenomenology of first-person agency. In: *Physicalism and Mental Causation* (ed. S. Walter and H.-D. Heckman). Imprint Academic.
- Loewer, B. (1996). Freedom from physics: quantum mechanics and free will. *Philosophical Topics* 24 (2): 91–112.
- Lowe, E.J. (2008). *Personal Agency: The Metaphysics of Mind and Action*. Oxford: Oxford University Press.
- Markosian, N. (1999). A compatibilist theory of agent causation. *Pacific Philosophical Quarterly* 80 (3): 257–277.
- Mele, A. (2003). *Motivation and Agency*. Oxford: Oxford University Press.
- Mele, A. (2006). *Free Will and Luck*. Oxford: Oxford University Press.
- Mele, A. (2014). Luck and free will. *Metaphilosophy* 45: 543–557.
- Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2004). The phenomenology of free will. *Journal of Consciousness Studies* 11 (7–8): 162–179.
- Nelkin, D.K. (2011). *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.
- O'Connor, T. (1995). Agent causation. In: *Agents, Causes, and Events* (ed. T. O'Connor), 173–200. Oxford: Oxford University Press.
- O'Connor, T. (2000). *Persons and Causes*. Oxford: Oxford University Press.
- O'Connor, T. (2005). Freedom with a human face. *Midwest Studies in Philosophy* 29 (1): 207–227.
- O'Connor, T. (2008). Agent-causal power. In: *Dispositions and Causes* (ed. T. Handfield), 189–214. Oxford: Oxford University Press.
- O'Connor, T. (2011). Agent-causal theories of freedom. In: *The Oxford Handbook of Free Will*, 2e (ed. R. Kane), 309–328. Oxford University Press.
- O'Connor, T. and Ross Churchill, J. (2004). Reasons explanation and agent control: in search of an integrated account. *Philosophical Topics* 32: 241–253.
- Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
- Vicens, L. (2016). Objective probabilities of free choice. *Res Philosophica* 93 (1): 125–135.
- Watson, G. (2003). *Free will*, 2e. Oxford: Oxford University Press.

Obligation and Moral Responsibility

ISHTIYAQUE HAJI

Appraisals of moral praiseworthiness and blameworthiness, or appraisals of moral responsibility, differ from those of moral permissibility, impermissibility, and obligation, or morally deontic appraisals. We can better appreciate these different sorts of normative assessment by exploring their pertinent similarities and differences. In section 1, I record a similarity: responsibility and obligation both presuppose freedom. I then motivate the view that obligation requires freedom to do otherwise. In section 2, I argue that a suitably amended objection to responsibility's requiring alternative possibilities, which invokes Frankfurt examples, is ineffective against the view that obligation requires alternative possibilities. In section 3, I offer considerations against the principle that blameworthiness is conceptually linked to impermissibility. In section 4, assuming that you have a moral obligation to do something only if you have the ability to do it, I develop a dilemma concerning this ability.

1 Freedom Requirements

In what follows, unless otherwise specified, "obligation" denotes overall and not *prima facie* moral obligation and "ought" in its occurrences in "ought statements" expresses overall moral obligation. The morally deontic "ought" is to be further distinguished from the morally ideal "ought," as when you claim that there ought to be no starving children, and other varieties of "ought" like the prudential or legal "ought." Instances of the canonical form of "ought" statements, such as, *at t1, Sid ought to keep her promise at t5*, capture the agent- and time-relativized dimensions of moral obligation. The first temporal index indicates the time of the obligation's acquisition and the second the time of its discharge. Frequently, in what follows one or both of these indices will be omitted when their omission is inconsequential.

Although there are dissenters (e.g., Schlossberger 1992), it is widely believed that responsibility requires freedom. This is the freedom to do otherwise in a venerable and long-standing tradition. As Harry Frankfurt (1969) so famously framed the relevant principle, only to argue

against it, persons are morally responsible for having done something only if they could have done otherwise (PAP).

Obligation, like responsibility, requires freedom; nothing can be obligatory for you unless you have obligation-relevant freedom concerning it. *Kant's Law* that "ought" implies "can" (and "ought not" implies "can refrain from") captures an element of this freedom. Although controversial, a powerful analysis of obligation supports *Kant's Law*. This analysis explains the concept of obligation in terms of the concepts of possibility and goodness. It says that if, at the time of choice, there is a complete life history – a complete way in which your life can thereafter unfold – in which you perform an action, and no better life history in which you don't perform this action, then it is obligatory for you to perform this action. More rigorously, assuming we can rank worlds according to how good they are or in terms of their deontic value,

MO: You morally ought, at t , to do A at t^* , if, and only if, there is a world, w , accessible to you at t in which you do A at t^* ; and it is not the case that you avoid doing A at t^* in any world as good as or better than w that is accessible to you at t . (See, e.g., Feldman 1986, ch.2)

Just what deontic value consists in, whether, for instance, it is to be identified with intrinsic value, need not detain us. Simplifying, *MO* entails that you ought to do something if, and only if, you do it in all the best worlds accessible to you; it's permissible for you to do something if, and only if, you do it in some of the best worlds accessible to you; and it's impermissible for you to do something if, and only, you don't do it in any of the best worlds accessible to you. Since worlds accessible to you are worlds you can make actual, *MO* validates both a version of "ought" implies "can" and "permissible" implies "can."¹

A modified account of *MO* that includes an alterability clause:

MO-Refined: You ought, at t , to do A at t^* , if, and only if, there is a world, w , accessible to you at t in which you do A at t^* ; it is not the case that you avoid doing A at t^* in any world as good as or better than w that is accessible to you at t ; and, there is a world, w^* , accessible to you at t at which you avoid doing A at t^* ,

entails that "ought" implies both "can" and "can refrain from." But what warrants this modification? The answer is that there is independent reason to believe that obligation, impermissibility, and permissibility require that you can do otherwise. Here, I focus on obligation and impermissibility.

Starting with impermissibility, and presupposing that you ought not to do something if, and only if, it is impermissible for you to do it (*Equivalence*), impermissibility requires freedom to do otherwise: If it's impermissible for you to do something, you ought not to do it; if you ought not to do something, you can refrain from doing it; therefore, if it's impermissible for you to do something, you can refrain from doing it.

Does "impermissible," just like "ought," imply "can" too? Opponents of this principle claim that some acts can be impermissible for you even if you can't do them. It suffices to show that this is false if we can show that no act can be impermissible for you, and so *prohibition skepticism* is true.

Determinism is the thesis that the conjunction of the laws of nature and the state of the universe at any time entails the state of the universe at all later times. One event deterministically causes a second if, and only if, the first causes the second, and given the laws of nature and the past, there is no chance that the first occurs without causing the second. An event indeterministically causes another if, and only if, the former causes the latter, and it is consistent with the laws of nature and the past that the former occurs and not have caused the

latter. To establish prohibition skepticism we show that any act, whether deterministically or indeterministically produced, cannot be impermissible for you.

On the first option, assuming that determinism precludes freedom to do otherwise, if you do *A*, and *A* is deterministically produced, then *A* is unalterable for you; it occurs in every world accessible to you.² Since you cannot avoid doing any act that's unalterable for you, and impermissibility presupposes avoidability, no act of yours that's deterministically produced is impermissible for you.

On the second option, either *A* is distally or proximally indeterministically produced. With the former, you acquire indeterministically some element, such as a belief, in an action's etiology, and this belief, together with other apt mental items or states, deterministically gives rise to this action. With the latter, indeterminism is located between, roughly, your reason states and your action.

With one exception, no distally produced act is impermissible for you because if you perform this action at time *t*, at *t*, it's unavoidable for you as prior springs of action deterministically produce it. The exception is a case where the indeterministically acquired element, say belief *b*, and your putative obligation, the obligation to do *B*, are acquired at the *same* time, *t**. Imagine that had you not acquired, at *t**, belief *b*, you would have acquired, at *t**, belief *a* instead, and had you acquired *a*, you would have had the putative obligation, at *t**, to do *A*. But neither *B* nor *A* can be obligatory for you because you have no control over the acquisition of belief *b* (or *a*); it's a matter of luck which you acquire at *t**; hence, your *B*-ing (or *A*-ing) at *t** is, itself, too luck-infused to be obligatory for you. The luck here is of the same sort described below in connection with proximally produced acts.

To show that no proximally indeterministically produced act is impermissible for you, begin with the following principle (a theorem of *MO*):

Impermissible/Obligation Possibility (IOP): If, at *t*, it is impermissible for you to do *A* at *t**, then, at *t* you can do something else, such as refraining from doing *A* at *t**, that's obligatory for you.

Now suppose that, at *t*, you do *A* at *t** – a proximally produced act that's putatively impermissible for you – in world *w*. Since *A* is indeterministically produced, there is a world, *w**, with exactly the same past up to *t*, and the same laws, in which, at *t*, you do otherwise than *A* at *t**; you do *B* instead. It's plausible to judge that, at *t*, your doing *B* at *t** in world *w**, or the cross-world difference between *w* and *w**, is a matter of obligation-undermining luck, and so *B* cannot be obligatory for you. This judgment may be supported by appealing to the following nonexhaustive list:

- (1) Since there is no appropriate causal connection between your reason states and your deciding at *t* to *B*, you have no obligation-relevant control in *B*-ing in *w**.³
- (2) As an indeterministic agent, it's false that you have some further power to influence causally which of your alternatives you realize, a power over and above the mere chance of acting differently and to exercise control in doing whatever you do.⁴
- (3) If one agent does one thing and another refrains from doing it, and nothing about the agents' powers, capacities, states of mind, moral character and the like explains this difference, then the difference just is a matter of luck.⁵

Given *IOP*, if *A* is impermissible for you (as presumed), then there is something you can do instead, *B*, that's obligatory for you. But *B* is *not* obligatory for you because it is too luck-infused. Hence, since *IOP* is violated, *A* is not impermissible for you. We may conclude that no proximally produced act is impermissible for you.

Since any action must fall into one of the classes of either unalterable or alterable acts that have been identified, and no act in any of these classes can be impermissible for anyone, it's false that some act that you cannot do can be impermissible for you. Thus, we are entitled to conclude that "impermissible" implies "can."

It's now easy to confirm that not only does "ought" imply "can," it also implies "can refrain from": if you ought to do something, then it's impermissible for you to do otherwise than this thing (from *Equivalence*). If it's impermissible for you to do otherwise than this thing, you can refrain from doing otherwise than this thing (from "impermissible" implies "can"). Therefore, if you ought to do something, then you can refrain from doing otherwise than this thing. Thus, if you ought to do something, then both you can do, and you can refrain from doing, it.⁶ As obligation and impermissibility require dual control – the ability both to do, and to refrain from doing, whatever is obligatory or impermissible for you, we are justified in shifting from *MO* to *MO-Refined*.

2 Determinism, Responsibility, and Obligation

Consider these arguments for the incompatibility of determinism with responsibility and obligation respectively:

No Responsibility

- (1R) If determinism is true, no one can do otherwise.
- (2R) If no one can do otherwise, no one is ever morally responsible for anything.
- (3R) Therefore, if determinism is true, no one is every morally responsible for anything.

No Obligation

- (1O) If determinism is true, no one can do otherwise.
- (2O) If no one can do otherwise, no one ever has a moral obligation to do anything.
- (3O) Therefore, if determinism is true, no one ever has a moral obligation to do anything.

Lines (1R) and (1O) are based on the admittedly contentious view that determinism rules out freedom to do otherwise. For our purposes we may grant these premises with one qualification, to be introduced later, having to do with the analysis of "can." Whereas (2R) rests on PAP, (2O) appeals to the view that obligation requires freedom to do otherwise.

I briefly comment on whether a popular objection to (2R), suitably modified, is an effectual objection to (2O). Many contend that Frankfurt examples undermine PAP and, so, undermine (2R). In Stage 1 in *Theft*, a Frankfurt case, although Jones could have refrained from stealing some pears, he is blameworthy for stealing them. In Stage 2, a "rerun" of Stage 1, something precludes Jones from doing anything incompatible with stealing but *without* in any way interfering in Jones's actually stealing as it turns out. A mind reader, Black, who can tell what Jones is about to do, will do nothing if he detects some reliable sign Jones displays that he, Jones, is about to steal, but will force Jones to steal if he discerns the reliable sign that Jones is about to refrain from stealing. But Jones proceeds exactly as before, so Black has no need to intercede. Since Jones in the absence of Black's intervention is morally blameworthy for stealing, and since in Stage 2 Jones behaves no differently, he is morally blameworthy for stealing here too, even though he could not have done otherwise.⁷ Hence, many contend, PAP is false.

Some might propose that if Frankfurt examples undermine PAP, they undermine *Kant's Law* too. Assume that in Stage 1 Jones ought not to steal the pears. Intuitively, it is also impermissible for Jones to steal the pears in Stage 2, even though he could not have refrained

from stealing them. As this is so, and the conjunction of *Equivalence* and *Kant's Law* entails that impermissibility requires avoidability, assuming *Equivalence* is unassailable, the culprit is *Kant's Law*. If, in turn, *Kant's Law* is false, we can block the derivation of "ought" implies "can refrain from" from *Kant's Law*, *Equivalence*, and "impermissible" implies "can."

However, this way of rejecting the principle that impermissibility (or obligation) requires avoidability is problematic. *Kant's Law* is pertinently like the following principle.

Blame/Control: You are blameworthy (or, more generally, morally responsible) for performing an action only if you can perform it.

This principle expresses the highly plausible view that blameworthiness (or, more broadly, responsibility) requires control. Notably, Frankfurt examples do *not* undermine this principle. The principle of alternate possibilities regarding blameworthiness:

PAP-Blame: You are morally blameworthy for having done something only if you could have done otherwise,

is a conjunction of *Blame/Control* and

Action: For any event and any agent, the event is the agent's action only if she has the ability to preclude it from occurring (Zimmerman 1996, p. 86).

Action provides the alleged link between blameworthiness and alternative possibilities, but Frankfurt examples undercut it.

If Frankfurt examples leave unscathed the principle that blameworthiness requires control (*Blame/Control*), they should leave unscathed the principle that obligation requires control (*Kant's Law*) or that impermissibility requires control:

Kant's Law/Impermissible: It is impermissible for you to do something only if you can do it.

The link between obligation and alternative possibilities is provided not by *Action* but by *Kant's Law/Impermissible* and *Equivalence*. Just as Frankfurt examples do not impugn the principle that impermissibility (or obligation) requires control, so they do not impugn *Equivalence*.

There is, then, this significant difference between *PAP-Blame* and the principle that obligation requires alternatives. Regarding the former, essential to the link between blameworthiness and alternative possibilities is principle *Action*, but Frankfurt examples undermine *Action*. Regarding the latter, essential to the link between obligation and alternative possibilities are *Kant's Law/Impermissible* and *Equivalence*. Frankfurt examples threaten neither of these principles. Minimally, if one grants that these examples leave untouched the principle that blameworthiness requires control, one should also grant that they leave untouched the principle that impermissibility (or obligation) requires control. Moreover, it is implausible to suppose that Frankfurt examples undercut *Equivalence*.

3 Blameworthiness and Impermissibility

The principle that blameworthiness presupposes impermissibility (*BRI*) challenges Frankfurt examples, such as *Theft*, in which the primary agent is putatively blameworthy.

BRI: Necessarily, you are morally blameworthy, for example, for an action, only if it is impermissible for you to do it (see, e.g., Widerker 1991, p. 223; Copp 1997; Fischer 2006, p. 218; Campbell 2011, pp. 33–34).

Recall, since impermissibility requires avoidability, in Stage 2 it isn't impermissible for Jones to steal the pears because, here, he cannot refrain from stealing them. But if *BRI* is true, contrary to what Frankfurt defenders propose, Jones is *not* blameworthy for stealing the pears in Stage 2.

BRI, however, is suspect. First, it has putative counterexamples (see, e.g., Zimmerman 1997; Haji 2002; Vranas 2007; Scanlon 2008; Capes 2012). Second, supererogatory acts pressure *BRI*. Elaborating briefly, supererogatory acts are permissible acts that have a special sort of value. If there are ranked permissible options, some such options being better than others, one way to conceptualize supererogatory acts, as Paul McNamara proposes, is as follows. An act is supererogatory for you if, and only if, it is optional for you to do, praiseworthy for you to do, not blameworthy for you to omit, and precluded by doing the least you can do; that is, precluded by your lowest ranked permissible option (see, e.g., McNamara 2011, p. 223).⁸ Suberogatory acts, are, roughly, the symmetric mirror counterparts of supererogatory ones; they are permissible acts with a special sort of disvalue. An act is suberogatory for you if, and only if, it is optional for you to do, blameworthy for you to do, not praiseworthy for you to omit, and precluded by your top-ranked permissible option or the maximum that you can do (McNamara 2011, p. 231). If there are suberogatory acts, *BRI* is false (Haji 1998, 2002, 2012).

Why think that there are suberogatory acts? A category much like that of the suberogatory has an ancient lineage. Prevalent in Islamic ethics is a fivefold model of the moral classification of acts: An act may be morally obligatory, recommended, permissible, discouraged, or forbidden. Recommended acts, akin to supererogatory ones, are permissible acts for the doing of which there is a reward – they are commendable – but for the neglect of which there is no punishment. The class of discouraged acts bears striking resemblance to that of suberogatory acts in a certain respect. Discouraged acts are permissible, and, although there is no punishment for their performance, their omission is rewarded.⁹ Roderick Chisholm (1963, p. 5) seems initially to have become aware of the suberogatory through its status as a symmetric logical counterpart of the supererogatory. If it's possible for you to discharge your obligations in a supererogatory way (in a better than minimal way), then it ought to be possible for you to discharge them in a minimal way too (see, e.g. McNamara (1996, p. 426)). Additionally, suberogatory acts have been invoked to illuminate theoretical puzzles, such as having an *owed favor* (Driver 1992; McNamara 2011), or why there is something morally untoward with an eighth successive abortion, even if none of the eight is impermissible (Driver 1992). Finally, an *MO*-like framework for obligation validates this category (McNamara 1996).

Third, I have also argued that considerations of luck undermine *BRI*. To outline the argument, something is a matter of luck, loosely, if it is beyond our control. For example, it is a matter of luck that Jones does no wrong in stealing the pears in Stage 2 because he had no control whatsoever over Black's behavior. We may now introduce the following principle:

Luck Principle: If your doing a certain sort of thing in a certain way, including freely doing it, is sufficient for being blameworthy for doing it in that way, then something with respect to which you had no personal control, cannot affect your degree of blameworthiness for doing the sort of thing you did in the way (including freely) in which you did it.

Frankfurt examples support the luck principle. In *Theft*, it isn't in Jones's power to influence any of Black's activities. With Black absent in Stage 1, assume that Jones is blameworthy, to some degree, for stealing the pears. If Black is present, as in Stage 2, Jones is just as blameworthy for stealing the pears in this stage as he is for stealing them in the former. Black's presence makes no difference to Jones's degree of blameworthiness because Jones has no control over Black's conduct.

The argument from luck against BRI may now be summarized thusly:

(L1) It is impermissible for Jones to steal the pears and he is blameworthy for stealing them in Stage 1.

(L2) *Luck Principle* (abridged): The degree to which you are blameworthy cannot be affected by what is not in your control.

(L3) Since the only difference between Stage 1 and Stage 2 is the presence of Black and, hence, something Jones has no control over, *Luck Principle* would be violated if Jones were blameworthy in Stage 1, but not in Stage 2.

(L4) Therefore, Jones is blameworthy in Stage 2 as well.

(L5) Jones's action is not impermissible in Stage 2, however, because impermissibility requires alternative possibilities, and Jones does not have any alternative possibilities in Stage 2.

(L6) In Stage 2, then, Jones is blameworthy for stealing the pears, although it is not impermissible for him to steal the pears.

(L7) Hence, blameworthiness does not require impermissibility.

4 A Dilemma: Does Obligation Require Weak or Strong Alternatives?

Addressing a left over issue with *No Obligation*, this argument seeks to convince us that since determinism effaces alternatives (10) that obligation presupposes (20), determinism and obligation are incompatible. The remaining issue concerns the sort of alternative that obligation requires. Weak alternatives are alternatives you can have even if determinism is true. More precisely, you have a weak alternative provided that, given a subset of facts about the past and the laws, you could have done otherwise. Strong alternatives are alternatives that are incompatible with determinism. If at t , you do A at t^* , in world, w , then at t , you have a strong alternative only if, there is some world, w^* , with exactly the same pre- t past and the same laws as w , in which at t , you don't do A at t^* ; either, at t , you refrain from doing A at t^* or do something else, B , at t^* . As determinism does not expunge weak alternatives, *No Obligation* is not sound if "alternative" in each of its premises refers to weak alternatives. If "alternatives," however, in these premises denotes strong alternatives, then we confront the complex issue of whether obligation requires strong alternatives. Here, my modest aim is to sketch and briefly evaluate a strategy supporting the view that obligation requires weak alternatives.

Libertarianism regarding responsibility says that determinism is incompatible with responsibility, and there are free actions for which some people are morally responsible. *Libertarianism regarding obligation* is the view that determinism is incompatible with obligation, and there are free actions that are obligatory for some people. Typically, libertarians (or others) regarding responsibility who believe that responsibility requires alternatives claim that these alternatives are strong. I confine discussion to two reasons that have been advanced to support this claim. This will help to illuminate the strategy to argue for obligation's requiring weak alternatives.

The first reason invokes ultimate origination. The underlying thought is that if we are to be morally responsible for our behavior, we must be the underwriters of this behavior in a way in which we could not be if determinism were true. For if all our behavior were deterministically caused, then there is a sense in which the facts of the distant past, in conjunction with the laws of nature, are the sources of this behavior. Derk Pereboom proposes that if you decide to perform an action solely because, unbeknownst to you, neuroscientists have manipulated you to decide in accordance with their surreptitious neural programming, then it appears that your decision is produced by a source over which you have no control. Hence, since you aren't the ultimate source of your decision, you are not responsible for that decision (see, e.g., Pereboom 2001, pp. 4, 43).

Robert Kane writes:

[W]e may manipulate ... [people] into doing what we want while making them feel that they have made up their own minds ... which is covert nonconstraining or CNC control. Cases of CNC control in larger settings are provided by examples of behavioral engineering such as we find in utopian works like Aldous Huxley's *Brave New World* or B. F. Skinner's *Walden Two* ... [M]ost people who look at *Walden Two* would say – and say rightly, in my opinion – that its citizens lack free will in a deeper sense than being able to do what they want and will what they want. In this deeper sense, their wills are not “their own” because they are not the original creators of their own ends or purposes (1996, p. 65, note omitted).

A decision is directly free only if it is free and its freedom does not derive from the freedom of other actions to which it is suitably related. Kane proposes that if a decision is directly free, then it is indeterministically and nondeviantly caused by your reason states. He says that an agent is the ultimate source of such indeterministically caused (free) decisions. Regarding directly free actions, the condition that the agent had strong alternatives – the agent could have done otherwise, given the same past and the laws – is necessary for the agent's being the ultimate originator of such actions. In other words, having strong alternatives is required for an agent to be the ultimate originator of at least some of her choices.

Suppose we grant that ultimate origination, as Kane conceives of it, is essential to responsibility, and, hence, responsibility requires that you have strong alternatives on those occasions when your actions are directly free. Should we think that obligation also presupposes that we be the ultimate originators of our actions? Hardly. To explain, imagine that you have fallen victim to manipulative psychosurgeons. Via directly stimulating your brain, they leave you with two and only two engineered in mutually exclusive choices: you can at t intentionally do A or B at t^* . Had you not fallen victim to these psychosurgeons, you would not have had either of these options. Nothing precludes it from being such that given the manipulation, you have only two alternatives, and no matter which you perform, either is indeterministically caused. Suppose, at t , you do A . Fashion the case so that you are not the ultimate originator, as Kane conceives of ultimate origination, of A . You fail to satisfy the following condition that Kane recommends. To be ultimately responsible for an action, you must be responsible for anything that is a sufficient cause or motive for the action's occurring (1996, p. 72). Still, we may suppose that, given these alternatives, in all the best worlds accessible to you at t , you do A at t^* . Why should the fact that you are not the ultimate originator of A preclude A from being morally obligatory for you? Given the manipulation, you make the world as good as you can by doing A . Or imagine that instead of doing A you deliberately do B . Why, in doing B , would you not have done wrong? Deontic assessments of obligation are *act-focused* (the act being the primary target of evaluation) unlike assessments of responsibility that are *agent-focused* (the agent being the principal object of evaluation). It's plausible to suppose that an agent is not

responsible for an action if she is not its ultimate source, partly but pivotally, because responsibility assessments are agent-focused. In contrast, if coercion need not preclude you from doing what is obligatory (or impermissible) for you, CNC manipulation – something incompatible with ultimate sourcehood – should not preclude this either. Again, this is partly but centrally because deontic assessments are act-focused: An act can instantiate the property of *being such that it is the act you perform in all the best worlds accessible to you at some time* even if you perform it only because you are a victim of CNC manipulation.

The moral here is that what may be a plausible reason, associated with ultimate origination, to think that responsibility requires that you have strong alternatives on at least some occasions, is *not* also a plausible reason to think that obligation requires that you have such alternatives.

Consider, next, a second reason why it may be thought that responsibility presupposes your having strong alternatives. An attractive idea is that you are morally responsible for your actions only if it is not the case that all the contributions you make to your actions are deterministic products of the state of the universe in the distant past and the laws. Indeterministic choice gives us *freedom from control by the past*. As Alfred Mele proposes:

The capacity to make and execute intelligent decisions that are indeterministically caused by their proximal causes gives agents who have it a measure of independence from the past ... Agents with this capacity can make intelligent contributions to their world of such a kind that it is false that their every thought and action is part of a deterministic causal chain (Mele 2017, p. 184).

Some may think that there is no essential difference between freedom from control by the past and ultimate origination. However, this would be a mistake as we will see.

Addressing the question of immediate interest, does *obligation* require freedom from control by the past? Even if determinism is true, we lack control over the coming to mind of many beliefs that, if they were to come to mind, would play a substantial causal role in producing various decisions we make. Consider, first, this sort of scenario featuring indeterminism: Imagine that Abe's decision at t_5 to do B (at t_5) is deterministically caused by, among other elements, a proximal best-from-his-point-of-view judgment concerning what to do. This judgment, in turn, was causally produced by an apt desire, and a belief, B_f , that indeterministically came to his mind at t_0 . Despite Abe's not having control over acquiring this belief, he has proximal control in deciding to B . Furthermore, he has freedom from control by the past in making his decision to B because B_f was indeterministically acquired: Given the same past and the laws, there is a possible world in which a different belief, A_f , comes to his mind at t_0 , and if he acquires A_f , he decides to A (rather than B) at t_5 . Consider, next, this deterministic scenario that mirrors the former indeterministic one in all respects save one: B_f deterministically comes to Abe's mind at t_0 . Underscoring a few things about these scenarios, first, in both Abe decides at t_5 to do B only because belief B_f came to his mind at some earlier time. Second, in neither scenario is it plausible to suppose that Abe is the ultimate originator of his decision to B , at least not in the way Kane understands the concept of *being an ultimate originator*, because Abe has no control over acquiring B_f at the time at which he did. Third, however, in the first but not the second scenario, Abe has freedom from control by the past. It is, thus, false that having freedom from control by the past amounts to nothing different than being an ultimate originator.

Suppose you concede that at t_3 it is *obligatory* for Abe to decide to B at t_5 in the indeterministic scenario. Why would there, consequently, be any reason to deny that Abe has this this obligation in the deterministic scenario as well? To elaborate, it is challenging to argue

that whether it is obligatory for Abe to do *B* turns on whether or not he indeterministically acquires *Bf* *provided* it is acknowledged that he ought to do *B* in the indeterministic scenario. Since Abe has no control in acquiring *Bf* in either scenario, if it were denied that Abe ought to do *B* in the deterministic scenario, it would seem that it is merely *how* Abe acquires *Bf* in either scenario that makes the difference to the moral status of his *B*-ing in each of the scenarios if you were to insist on such a difference. But it is hard to see how this *can* make such a difference. Why should *Bf*'s being indeterministically acquired as opposed to being deterministically acquired, when in either scenario Abe has no control over acquiring *Bf*, translate into a difference in the moral status of Abe's *B*-ing?

In sum, some may venture that obligation requires strong alternatives because, first, responsibility requires such alternatives, and, second, the reasons advanced for the view that responsibility requires strong alternatives are, with suitable modification, *also* reasons to believe that obligation requires such alternatives. I suspect, however, that this strategy is not promising, and this may, in turn, incline you to the view that obligation requires weak alternatives.

You may object that if the "can" of obligation is some relatively weak compatibilist "can," then the argument previously given for "impermissible" implies "can" is not sound, as this argument presupposes that if determinism is true, you can't, in the sense of some incompatibilist "strong can't," do otherwise. In reply, I offer three comments. First, the previous argument has the virtue of at least raising a *prima facie* problem for the compatibility of impermissibility (or obligation) and determinism that invites a response. Second, you may respond by insisting that the "can" of obligation is weak, but if you do, you will require an argument for the "can's" being weak. Third, and relatedly, the relevant verdict isn't yet in. It may well be that obligation requires that we have strong alternatives.

You may, of course, still wonder about what more may be said in favor of obligation's requiring weak alternatives. Here, I end with a dilemma of sorts. Among weak views, some are weaker than others. To facilitate discussion, assume that "can" expresses ability, this ability is weak and, hence, as we may say, this "can" is weak, and an obligation to do something, as previously argued, requires the ability to do, and to refrain from doing, it. In addition, I will conduct the ensuing discussion in terms of weak and strong "can." For you to have a weak (or strong) alternative to doing *A* is for you to be weakly (or strongly) able to refrain from doing *A*. Roughly, at *t*, you strong-can do *A* only if your doing *A* at *t* is consistent with all pre-*t* facts of the past and the laws. If you strong-can, at *t*, do *A*, and determinism is true, you strong-cannot, at *t*, refrain from doing *A*. Roughly, you weak-can, at *t*, do *A* only if your doing *A* at *t* is not precluded by a *relevant* subset of pre-*t* facts of the past and the laws; at *t*, your doing *A* is consistent with this subset of facts and the past. Now we may reframe the primary question in this way: How is the weak "can" that you may theorize obligation presupposes to be analyzed?

Let's start with Dana Nelkin's proposal, energized partly, it seems, by her reflection on Frankfurt examples. Accepting *Kant's Law* and *Equivalence*, Nelkin concurs that if it is impermissible for you to do something, you can refrain from doing it. But she also endorses *BRI* and accepts the consequence that blameworthiness requires avoidability. Finally, acknowledging that Jones is blameworthy for stealing the pears in Stage 2 in *Theft*, Nelkin inherits the burden of explaining how Jones could, indeed, have done otherwise in Stage 2 of his Frankfurt predicament if it is true, as she believes, that it is indeed impermissible for Jones in Stage 2 to steal the pears. She explains:

What is needed to remove one's ability to do something in the relevant sense (call it the "interference-free capacity") is either the removal of the capacities, talents, skills, and so on (the presence of which is not in dispute) or the actual interference with or prevention of the exercise of those capacities. The fact that...[Black's] device *would* interfere or prevent such an exercise in counterfactual circumstances does not entail *actual* interference or prevention (2011, pp. 66–67).

Nelkin's suggestion regarding how the weak "can" of obligation is to be understood seems to amount to the following:

NC: At t you can do A at t^* if, and only if, at t (i) you have the talent, skills, knowledge, and so on that are necessary for doing A at t^* , in brief, you have the capacity at t to do A at t^* , and (ii) nothing interferes with your exercising this capacity at t^* .

NC encodes a weak sense of "can" because it implies that if you can, at t , do A at t^* , then at t your doing A at t^* is consistent with a *subset* of pertinent contingent facts of the past – among other facts, facts having to do with your germane skills, talents, and so forth – and the laws of nature.

If "can" expresses the ability (and opportunity) that obligation demands, and NC captures a necessary condition of your having this ability, then *Kant's Law* may be reconceived in this way:

New-Kant's Law: You ought, at t , to do A at t^* only if you have the capacity at t to do A at t^* and nothing interferes with your exercising this capacity at t .

NC, however, is too weak to curry favor with many proponents of *Kant's Law*. To appreciate how weak it is, imagine that in *Scene-1*, a child is drowning. You can easily rescue her. Assume you have an obligation to do so. *Scene-2* is just like the former save that, unbeknownst to you, there is a diver in the water between you and the child who will kill you if you enter the water. Intuitively, as of the time you have not entered the water and the diver is keeping watch, you have no obligation to save the child (although you may *think* you have such an obligation). In contrast, presumably, proponents of *New-Kant's Law* will claim that, at this time, you do have an obligation to save the child: Although you cannot, at this time, bring it about that you save the child, you NC-can intentionally save the child; nothing – certainly not the diver – as of the time you have not yet entered the water – interferes with your capacity to rescue the child.

To evade results of this sort, the obvious next move is to attempt to strengthen NC by *expanding* the set of facts about the past so that your doing A at t is compossible with the conjunction of *these* facts and the laws of nature. Regarding the example with the hidden diver, you may suggest that not only should we take into account facts about your germane talents, rescuing skills, and so forth but also relevant facts about the particular situation in which you find yourself, such as, the presence of the diver in the water. But this sort of expansion generates a problem. How is this subset of facts, partly in terms of which the "can" that obligation presupposes is to be analyzed, to be broadened to exclude cases (like the one with the hidden diver) in which it looks as though, intuitively, you have no obligation to do something (like saving the child)? Proponents of strong "can," exploiting this worry, may now play their trump card: If the strategy to amend weak "can" (such as "NC-can") to avoid putative counterexamples to your having an obligation to do something is to strengthen the "can" by expanding the subset of relevant facts, there is no non-arbitrary way to adjust this subset so that it contains all and only those truths pertinent to your ability to do whatever it is that you (supposedly) have an obligation to do. Why, then, not go all the way to strong "can"? Why not settle for the following: You ought, at t , to do A at t^* only if your doing A at t is (roughly) consistent with the conjunction of *all* the pre- t facts of the past and the laws of nature? The pertinent dilemma regarding the "can" of obligation should now be clear: Either this "can" is weak or strong. If it's weak, it runs afoul of counterexamples. Since there seems to be no non-arbitrary way to strengthen "can" to avoid putative counterexamples without going all the way to strong

“can,” the “can” is strong. But there appear to be no conclusive reasons to believe that the “can” of obligation is strong. For example, reasons to think that responsibility presupposes strong “can” are not also reasons to believe that obligation, too, presupposes strong “can.”¹⁰

Notes

- 1 I say “a version” because, for instance, much more needs to be said about the interpretation of “can”.
- 2 I owe the following to Joe Campbell whom I thank. Van Inwagen (2015, p. 19) says *p* is humanly unalterable if and only if “*p* and nothing that any human being is or ever has been able to do is such that if someone were to do it, that person’s action might (could possibly) result in its not being the case that *p*.” My use of “alterable” is different and much closer to the following also provided by van Inwagen (2015, p. 19): “I mistakenly supposed that the only way in which it could be that one had no choice about the truth-value of a proposition would be for the truth-value of that proposition to be in some way so firmly “fixed” that one was unable to change it.” Van Inwagen continues: “I did not see that that there is another way for one to have no choice about the truth-value of a proposition: for that truth-value to be a mere matter of chance” (van Inwagen 2015, p. 19).
- 3 See e.g. Haji (2016, pp. 81–317).
- 4 See e.g. Clarke (2003, p. 96) and Pereboom (2014, pp. 31–39).
- 5 Mele (1999, 280). See also Mele (2013) in which he formulates the problem of present luck independently of appealing to any concerns of explanation.
- 6 Elsewhere (Haji 2016), I have argued that permissibility, too, requires that one have free will.
- 7 See, e.g., Frankfurt 1969; Kane 1996; Fischer 2006; Mele 2006.
- 8 See also Heyd 1982 and Mellema 1991, pp. 17, 125–129.
- 9 Kevin Reinhart reports that “It is noteworthy that Anasri [1972] finds that the five-fold system is implied in texts which predate the formal development of the system. It is reasonably clear, in any case, from the grammatical forms used (passive participle) that most of the terminology of the five-fold system is extra-Quaranic” (1983, p. 201, n.21).
- 10 This paper was completed during my tenure of a 2017–2021 Social Sciences and Humanities Research Council of Canada grant. I am most grateful to this granting agency for its support.

Bibliography

- Ansari, Z.I. (1972). Islamic juristic terminology before Shafii: a semantic analysis with special reference to Kufa. *Arabica* 19: 255–300.
- Campbell, J.K. (2011). *Free Will*. Oxford: Polity Press.
- Capes, J. (2012). Blameworthiness without wrongdoing. *Pacific Philosophical Quarterly* 93: 417–437.
- Chisholm, R. (1963). Superogation and offence: a conceptual scheme for ethics. *Ratio* 5: 1–14.
- Clarke, R. (2003). *Libertarian Accounts of Free Will*. New York: Oxford University.
- Copp, D. (1997). Defending the principle of alternate possibilities: blameworthiness and moral responsibility. *Nous* 31: 441–456.
- Driver, J. (1992). The suberogatory. *Australasian Journal of Philosophy* 70: 286–295.
- Feldman, F. (1986). *Doing The Best We Can*. Dordrecht: D. Reidel Publishing Company.
- Fischer, J.M. (2006). *My Way: Essays on Moral Responsibility*. New York: Oxford University Press.
- Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66: 829–839.
- Haji, I. (1998). *Moral Appraisability: Puzzles, Proposals, and Perplexities*. New York: Oxford University Press.
- Haji, I. (2002). *Deontic Morality and Control*. Cambridge: Cambridge University Press.

- Haji, I. (2012). *Reason's Debt to Freedom*. New York: Oxford University Press.
- Haji, I. (2016). *Luck's Mischief: Obligation and Blameworthiness on a Thread*. New York: Oxford University Press.
- Heyd, D. (1982). *Supererogation*. Cambridge: Cambridge University Press.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.
- McNamara, P. (1996). Making room for going beyond the call. *Mind* 105: 415–450.
- McNamara, P. (2011). Supererogation, inside and out: toward an adequate scheme for common-sense morality. In: *Oxford Studies in Normative Ethics, Volume 1* (ed. M. Timmons), 202–235. New York: Oxford University Press.
- Mele, A. (1999). Ultimate responsibility and dumb luck. *Social Philosophy and Policy* 16: 274–293.
- Mele, A. (2006). *Free Will and Luck*. New York: Oxford University Press.
- Mele, A. (2013). Moral responsibility and the continuation problem. *Philosophical Studies* 162: 237–255.
- Mele, A. (2017). *Aspects of Agency: Decisions, Abilities, Explanations, and Free Will*. New York: Oxford University Press.
- Mellema, G. (1991). *Beyond the Call of Duty: Supererogation, Obligation, and Offence*. Albany, NY: State University of New York Press.
- Nelkin, D.K. (2011). *Making Sense of Freedom and Responsibility*. New York: Oxford University Press.
- Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.
- Reinhart, K. (1983). Islamic law as Islamic ethics. *Journal of Religious Ethics* 11: 186–203.
- Scanlon, T.M. (2008). *Moral Dimensions*. Princeton, NJ: Princeton University Press.
- Schlossberger, E. (1992). *Moral Responsibility and Persons*. Philadelphia, PA: Temple University Press.
- van Inwagen, P. (2015). Some thoughts on an essay on free will. *Harvard Review of Philosophy* 22: 16–30.
- Vranas, P.B.M. (2007). I ought, therefore I can. *Philosophical Studies* 136: 167–216.
- Widerker, D. (1991). Frankfurt on “ought implies can” and alternative possibilities. *Analysis* 51: 222–224.
- Zimmerman, M.J. (1996). *The Concept of Moral Obligation*. Cambridge: Cambridge University Press.
- Zimmerman, M.J. (1997). A plea for accuses. *American Philosophical Quarterly* 34: 229–243.

Perfect Freedom

MARILYN MCCORD ADAMS

1 Varieties of Freedom

Freedom is usually seen as an excellence in human agency. Along with reason or intelligence, freedom is what distinguishes human beings from the beasts. That is enough to make understanding freedom important. Unsurprisingly, many notions of individual human freedom have been defended, critiqued, and rejected on philosophical and empirical grounds. Theories of freedom fall into families: compatibilism vs incompatibilism vs determinism vs free-will skepticism. Incompatibilisms split into source- vs power-for-opposites incompatibilism, and again into virtue-incompatibilism vs liberty of indifference. Because philosophical theologians and philosophers of religion are *philosophers*, they participate in these debates and feel the pull of purely philosophical motivations favoring some conceptions over others. But their task is complicated by the need to test accounts of freedom for religious or theological adequacy. What I want to explore here is the effect of such religious or theological desiderata on the way Christian philosophers understand human freedom.

The biblical rubric for what is distinctive about human nature is that human beings are made in God's image and likeness (Gen 1:26). Influenced by neo-platonism, patristic and medieval philosophical theology take for granted that all creatable natures bear some resemblance to the Divine essence. Many (including Anselm, Bonaventure, Aquinas, and Henry of Ghent) would say that to be is either to be God or to be a way of imitating God.¹ Since God is infinite and creatures are finite, no one creatable kind is – by itself – an adequate reflection of Divine perfection, but each kind highlights certain aspects. Likewise, Divine agency is excellent in many ways. At stake for philosophical theologians and philosophers of religion is which such excellences characterize human agency, and which (if any) constitute human freedom. For the discussion that follows, it will prove helpful to overview four.

1.1 Self-Determination

All agree that self-determination (independence of external causes) is a metaphysical excellence. Since God is utterly self-determined (with respect to Divine Being, Excellence, and causal activity), self-determination would be one candidate for the way the best creatable

A Companion to Free Will, First Edition. Edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White.
© 2023 John Wiley & Sons, Inc. Published 2023 by John Wiley & Sons, Inc.

natures imitate God. But it is obviously metaphysically impossible for any creature to be utterly self-determined: all creatures depend on God, not only for their existence and excellence (so long as they have them), but also for their activity (insofar as God generally concurs with the actions of any and all creatures).² Accordingly, some philosophical theologians seek some core within human agency that is within our power to control. Perhaps influenced by the Stoics, Abelard identifies it as the agent's power to consent or withhold consent.³ Augustine and Anselm, Scotus and Ockham locate it in the will. Augustine, Scotus, and Ockham identify this power of self-determination with the will's freedom.⁴

1.2 Power for Opposites

Power-for-opposites was not universally regarded as an unqualified excellence. After all, God is self-determined in loving the Divine essence, but God has no power for opposites: to love or not to love Godself. Likewise, tradition has it that God the Father and God the Son "breathe out" their love for one another and produce the Holy Spirit.⁵ But they do so necessarily; it is not within their power to do otherwise. God is surely self-determined, but does God in eternity also have a power for opposites when it comes to creating or not creating, to creating this world as opposed to some other? Anselm's *Monologion* does not make this clear.⁶ Abelard seems to argue that Divine Justice would mean that God can do only what God actually does.⁷ By contrast, Aquinas,⁸ Scotus,⁹ and Ockham¹⁰ all insist that it is eternally within God's power not to create or to create something else. And they seem to be moved not only by the debatable idea that having options (as opposed to creating and creating what God creates by the necessity of God's nature) would increase God's providential control, but also by the less controversial notion that God is an agent that acts for and chooses among available reasons. Where congeries of finite creatures are concerned, there are always reasons to create as well as reasons not to create them.

Already Anselm argued that at least temporary power-for-opposites would be necessary if any self-determination is to be found in creatures. For creatures are metaphysically derivative: they get their being, their natures, and their nature-constituting powers from God. Anselm inferred that no action that was necessarily consequent upon such natural powers would be self-determined. For rational creatures to be somehow self-determined in their will-acts, God has to set up a situation in which they have a power for opposites, which allows them to self-determine their orientation to (or defection from) justice.¹¹

1.3 Focal Sourcing

Because creatures cannot be utterly self-determined, some writers shifted attention from being utterly independent and the ultimate causal source of choices and actions, to being (what I will call) their focal source. Agents trivially *own* actions when those actions involve the exercise or actualization of their causal capacities. Writers like Aristotle insist that agents own actions in such a way that they are imputable to the agent for praise or blame, when the proximate causes of the choice/action find distinctive integration in the agent's psyche. Thus, Aristotle says that an action is voluntary for an agent when it issues from that agent's practical deliberation and choice. The functional norm for voluntary action involves reason and desires interacting in particular ways. When the agent's faculties are immature or defective, the choice/action might not be imputable. Likewise, when the agent is invincibly ignorant of morally relevant information, or unduly influenced by passion so that its practical reasoning is obstructed, or when passion is so strong that reason is bound, the choice/action fails to be voluntary. So one can distinguish voluntary from involuntary choices/actions on the basis of

whether the agent is the focal source, whether the causes of the actions are integrated in the agent's psyche in the proper way and there is no malfunctioning. Aristotle says that the agent is "lord" because of the focal source that controls his voluntary choices/actions. Aquinas will say that such voluntary actions are free.¹²

1.4 *Stable Orientation to the Good*

Augustine equivocates between identifying freedom with power for opposites (in *On Free Choice of Will*) and alternatively with stable orientation to the Good.¹³ The latter freedom involves power successfully to resist any causes that might work to turn the agent away from willing the Good, and the secure possession of cognitive and motivational states that leave the agent no reason to turn away. Anselm, trying to find a definition of "freedom" that univocally applies to all rational or intelligent agents – human beings in this life who have not yet stabilized, and God and the good angels whose orientation to the Good is unshakeable – identifies freedom with power to uphold justice for its own sake.¹⁴ To be sure, most medievals (Ockham excepted¹⁵) think that goodness is the proper object of will-power, which makes it metaphysically necessary that the will is only able to will under the guise of the good (insofar as it seems good) and so is stable in that orientation by metaphysical necessity. What Augustine and Anselm have in mind, however, is not that you can't choose dill-pickle ice cream without seeing some good in it, but stability in loving God above all and for God's own sake, and loving neighbor for God's sake: stability that has no reason to will the contrary. For cradle-to-grave human beings, such freedom is salvation's eschatological goal.

2 Freedom as the Source of Evil

Among analytic Christian philosophers, the majority report is that incompatibilist free will, understood as a self-determining power for opposites, is *essential* to human agency and so something that characterizes the human race throughout the ups and downs of its salvation-historical career. For many, the principal *theological* motive is that they see no better way to solve the problem of evil. Their twin contentions are that (1) that the excellence of free creatures is necessary to any world good enough for God to create; and (2) that free creatures in the universe bring along contingencies that God cannot control.

2.1 *The Classical Version*

In the West, this approach to the problem of evil finds its *locus classicus* in Augustine's *On Free Choice of Will*, Books I and III. Whether because God is above reproach or because God is perfectly good by nature, it would be impious not only to blame God for evil, but also to say that God is the cause of evil or sin. What is needed is some kind of created agency that can be the somehow independent source of evil. Incompatibilist freedom, understood as a self-determining power for opposites, seems to them to fit that bill. God creates free creatures, most notably, Adam and Eve in Paradise. But they misuse their freedom thereby introducing the evil of sin. They are agent-sources of evil who could have done otherwise. Appeal is made to principles in the neighborhood of Doing-Allowing and New-Intervening-Agent to shift responsibility for the origin of evil off God's shoulders onto the incompatibilist free created agents. God is not to blame. Creatures are. Evils in the universe that cannot be traced to human causality are explained as the natural and/or punitive consequences of sin.¹⁶

2.2 *Plantinga's Version*

Plantinga's famous Free Will Defense¹⁷ makes no use of Doing-Allowing or New-Intervening-Agent principles. All the same, he recognizes the presumptive weight of J.L. Mackie's charge: that a perfectly good God would want, if possible, to create a world free from evils.¹⁸ Plantinga's reply is that it is a distinctive feature of incompatibilist-free creatures that they bring along with them contingent properties that even an omnipotent God cannot control. For Plantinga (re-inventing Molinism), these include properties about what such creatures would do if actualized in each and every circumstance in which they could exist. Plantinga argues: incompatibilist free creatures are necessary to an excellent enough world; yet it is logically possible that each incompatibilist free creature would go wrong at least once, no matter what the circumstances in which it was created. Plantinga does not distinguish source-incompatibilism from power-for-opposites incompatibilism. Perhaps he takes both for granted. What is clear is that – on Plantinga's theory – if X were created in C, X would be the source of X's sinful action. But the merely possible X is not the source of X's being such that X would sin in C if actual in C. Neither God nor the merely possible creatures are. Such properties befall the essences of incompatibilist free creatures, considered as merely possible, like fates.

2.3 *Distinctive Focus*

Beginning with the problem of evil drives philosophical theologians and philosophers of religion to posit some created agency that can be the originative source of evil and to identify freedom with some relevant feature(s) of that. On the Classical View, this means looking to human agency at its beginning (with Adam and Eve in Paradise). For Plantinga, the focus widens to cradle-to-grave human careers. Some combine such free-will approaches to the source of evil with a soul-making theodicy.¹⁹ Humans are made in God's *image* insofar as they are created with source-incompatibilist freedom. Our *ante-mortem* assignment is to exercise our source-incompatibilist free will in such a way as to build character, to respond to Divine offers of grace that enable right choices and virtuous habit formation, and thereby to grow more and more into the *likeness* of God. For the Classical View, what is primarily at stake between God and humans is Divine judgment of our imputable performances. In medieval philosophical theology, one of the most entrenched arguments for bodily resurrection and against Averroës's claim that there is only one agent and one possible intellect for the whole human race is that human individuals must all stand before the great judgment seat of Christ and be held to account for what they/we have done!

2.4 *Counting the Costs*

Suppose incompatibilist freedom, a self-determining power for opposites, could help to solve the problem of evil. The supposition that cradle-to-grave humans have it carries significant theological costs.

2.4.1 *Divine Foreknowledge*

Tradition maintains that from eternity God has certain and determinate knowledge of the entire course of created world history. But future incompatibilist free choices are contingent. Aristotle argued that future contingents as not yet ontologically determinate are not yet anything to know. Not even God can know what is as yet unknowable any more than God or anyone else can know what is false.

2.4.2 *Predestination*

Tradition holds that from eternity God has predestined some to glory. Moreover, Divine predestination is certain: “God predestines Peter to glory & Peter never enters into glory” is a contradiction. Yet, tradition also lays it down that, according to God’s contingent policy, only those who persevere to the end enter into glory, and persevering to the end is a function of Peter’s incompatibilist freely conforming his choices and actions to God’s will. If future contingents are ontologically indeterminate, it would seem that at any point prior to Peter’s death it is still ontologically indeterminate whether Peter will persevere to the end and so whether he will be eligible for heaven. But Divine predestination from eternity should make it eternally determinate.

2.4.3 *Providence*

According to tradition, from eternity God exercises meticulous providential control over created world history. Yet, how is that possible if human beings have and exercise source-incompatibilist free will?

2.4.4 *Proffered Solutions*

Scotus and Ockham both insist that – despite any impairments due to Adam’s fall – cradle-to-grave the human will is a self-determining power for opposites. If Scotus struggles unsuccessfully to explain how God’s eternal knowledge of created choices is compatible with their freedom,²⁰ Ockham’s ingenious strategy is to take the “fore” out of foreknowledge and the “pre” out of predestination. Rejecting Aristotle’s assumption – that something is knowable at a time only if it is ontologically determinate at that time – Ockham maintains that it is enough for the item in question to be ontologically determinate at some time or other. Since future contingents will become ontologically determinate one way or the other in the future, propositions about future contingents can be determinately true (or determinately false) after all. Ockham proceeds to explain that “God foreknew from eternity that Peter will persevere to the end” is logically equivalent to “Peter will persevere to the end,” and “from eternity, God predestined Peter” to “Peter will persevere to the end.” Although “God foreknew from eternity that Peter will persevere to the end” and “from eternity, God predestined Peter” are verbally past-tense, they are in reality in part about the future. Ockham makes the intentional direction of God’s knowledge and of God’s predestining will just as ontologically indeterminate as Peter’s future free choice is.²¹

Contemporary sponsors of cradle-to-grave incompatibilist freedom have mostly chosen between two other approaches. Molinism²² assigns God “middle” knowledge of each and every possible incompatibilist-free creature: i.e., knowledge for each creature X and each circumstance C in which X could exist, whether or not X would sin in C. God cannot causally determine creatures’ source-incompatibilist free choices. But God can use God’s middle knowledge to exercise meticulous providence by coordinating what creatures would do in which circumstances to create combinations that yield the outcomes that God all-things-considered wants. Thus, middle knowledge promises to reconcile incompatibilist human freedom with God’s meticulous providential control, the certainty of Divine predestination, and God’s eternal knowledge of the whole of created world history.

Others question the philosophical coherence of Molinism. In particular, they doubt whether there are any such facts about a merely possible creature X as what X would incompatibilist freely do if actual in circumstance C. These sponsors of incompatibilist freedom bite the bullet and embrace “Open Theism,”²³ which admits that Aristotle was right: future contingents are not yet ontologically determinate and so not yet knowable. God does not have

certain and determinate fore-/eternal knowledge of future contingents. Neither is God in a position to exercise meticulous providence over the whole course of world history, including created source-incompatibilist-free choices. Even Divine omnipotence and God's superior grasp of natural necessities and probabilities will not prevent God from having to grapple with a partially open future!

2.4.5 *Post-Mortem Stability?*

This gestures towards a further problem (explicitly raised in the middle ages): *how can the drama of salvation permanently resolve into a happy ending if the saints still have the liberty of indifference in heaven? What assurance do we have that history will not repeat itself with Adam's race freely and deliberately falling again?* Contemporary analytic philosophical theologians who combine their source-incompatibilism with a soul-making theodicy are confident of their answer. Over the course of our lives, our source-incompatibilist free choices build up habits of character, which narrow the range of reasons for action that appeal to us. The wicked get to the point that they can see no reason for repentance, while the elect become so virtuous that they can no longer find any reason to turn away from God. By contrast, Scotus and Ockham insisted that virtues (vices) only incline; they do not take away our self-determining power for opposites by making vicious or sinful (virtuous) choices impossible.²⁴ No, Ockham insists, God makes the sure happiness of the elect and the unavoidable misery of the damned permanent by acting as the total efficient cause to produce a wholehearted love of God in the former and a hatred of punishment and will for happiness in the latter. The damned are prevented from repenting either by God's acting as a total efficient cause of their act of hating God or by God's refusing to concur with any acts of loving God on their part.²⁵

2.5 *Futile Appeal*

If incompatibilist freedom is posited in the first place to solve the problem of evil, it is appropriate to ask whether it really helps. My own verdict is negative. The Molinism that underlies Plantinga's approach is, as David Lewis wryly implies, at least as philosophically vexed as the problem of evil with which he started.²⁶ Classical appeals to principles such as Doing-Allowing and New-Intervening Agent are of no avail, because they apply only when the agents in question are near-enough peers. The agency of toddlers does nothing to distance supervising adults from responsibility for any harms the children cause. The "size-gap" between infinite and finite agencies would prevent even incompatibilist free creatures from standing as near-enough peers of Divine agency to shift responsibility for the origin of evil. The buck still stops with God!²⁷

3 Freedom as Perichoresis²⁸

I have just examined one strand of Western Christian tradition that identifies being made in God's image with being rational creatures who act through reason and free choice of will. Source-incompatibilism was supposed to give humans enough independence as agents to be sources of evil. Appended soul-making theodicies make the imputable performance of incompatibilist freedom a salient factor in the contrasting eternal destinies of human beings. According to that strand of the tradition, the damned deserve hell because they have brought it on themselves. And – while nothing human beings can be or do would make them intrinsically worthy of heaven – still the elects' track-record of incompatibilist free choices and responses

to Divine grace builds into them a character that somehow fits them for heaven. I turn now to an alternative approach that relocates the image of God in our being persons and reidentifies perfect freedom, not as an initial or lifelong possession, but as an eschatological goal.

3.1 *Personhood as Subjectivity*

Of course, tradition is right to insist that personal agency is intelligent voluntary agency. But we get closer to the essence of what it is to be a person when we say that *to be personal is to have a subjective point of view*. Here several points are noteworthy. First, because you are personal, there is such a thing as what it is like to be you (the way – *pace* panpsychism – there is no such thing as what it is like to be a rock or a chair, although there may well be such a thing as what it is like to be a bat).²⁹ Fully functioning persons are *self-conscious*, so that there is such a thing as what it is like *for you* to be you. Fully functioning persons also are able to recognize other human beings as selves with their own subjectivities, taking their own point of view, seeing and valuing, relating to the world and others in their own distinctive ways.

Second, *personhood is essentially relational*. The subjectivity of persons is evoked by and given content in relation to other persons. According to Christian theology, Godhead is paradigm personhood. Whether or not the Divine persons are metaphysically constituted by their relations to the other Divine persons (as the Western majority-report maintains), the subjective point of view of each of the Divine persons is shaped by its relations with the other two. Personal subjectivity does not arise or thrive in isolation.

Third, *human personhood is naturally developmental*. We begin life as packages of protoplasm organized in ways complex enough to host personal life, and so begin full of personal potential. But for us that potential requires to be evoked, evolved, and nurtured in a personal surround. What is evoked and evolved is our subjectivity, our subjective sense of ourselves as selves, our recognition of the people around us as other selves, the models in terms of which we structure ourselves and our interactions in relation to the world and to other people. Developmental psychologists have drawn a variety of maps charting characteristic ways in which our subjective point of view develops, how we come up with a model of who we are and who others are and how to relate to one another and the other things around us; how we stabilize and operate out of it for a time and a season, until our cognitive and affective capacities grow and/or in the face of traumatic experiences that root-and-branch challenge our way of being in the world so that the old model goes to smash and we are forced to gather up the pieces and evolve another.

Fourth, *human development is essentially interactive*, always evoking and forming *by involving* our agency, which goes through many stages to reach maturity. How we participate makes a big difference, but our interactive partners and environmental circumstances also make a big difference. Development is easy to arrest, to twist, to reverse, and to undo. Human personality is especially fragile at the beginning, not only in infancy but also in early childhood stages, where reactive patterns formed on the basis of the child's limited cognitive skills and affective repertoire become entrenched. Neglectful and abusive caretakers can do lasting damage. Traumatic experiences which swamp coping capacities can damage even adult agency (think of soldiers with PTSD or civilians of all ages in war-torn areas). Even in the best circumstances, we are shaped and formed by interaction with neurotic caretakers and teachers, and – normally and naturally – we adapt complementary neurotic strategies for coping with them.

Fifth, *developmental models of human agency are teleological*. To be sure, there are norms for each stage, skills and competencies to be acquired, attitudinal postures to be struck.

Age-appropriate practices of positive and negative reinforcement loom large in personal formation. Normally and naturally, when things go well, each new stage builds on while recontextualizing the achievements of the earlier.³⁰ But the whole process aims at the dynamic organization achieved when functional maturity is reached. To whatever extent earlier developmental stages involved the agent's initiatives, *perfect freedom would be expected only when the agent is fully mature.*

3.2 Maps and Norms

A variety of scales are offered to map and norm the individual's progress. According to psychoanalytic versions, the infant psyche is a booming buzzing confusion of inputs and impulses, but by the age of three months or so has the cognitive capacity to differentiate and center its psychic field on a human face. A few months later, its cognitive skills progress enough to recognize that the face goes away, and so cannot be relied upon to hold the self together. The infant differentiates itself from the face and begins a long process of ego development over the course of which the ego deploys a variety of self-management strategies that organize, structure, and restructure the personality. Therapy assists the ego in consciously identifying, sorting, and discarding dysfunctional defenses and – when successful – brings the individual to the goal of rational self-government.³¹

Lawrence Kohlberg³² charts the evolving structures of moral functioning, from punishment/reward to instrumental hedonism, onto local peer-group (tribalism) and then societal orientation, and finally to principled self-government according to universal and universalizing ethical principles (à la Kant). Robert Kegan³³ envisions the human individual structuring and restructuring its personality through sequences of stages in which the self is first embedded in a matrix – maternal, familial, peer group, social institutions – and then disembeds, then re-embeds, culminating in differentiated autonomy. Kegan contrasts the earlier self-constructions that identify and fuse with authority figures or social groups from the goal of autonomous differentiated individuals entering into intimacy with others differentiated from themselves.

Taking a page from Kant, these developmental psychologists appear to agree with many philosophers: the freedom that is a good-making feature in human being is differentiated autonomy, that is and recognized itself to be functionally independent when it comes to its beliefs and choices. Differentiated autonomy does not “follow the leader” no matter what – whether that leader is parent, teacher, or mentor. Neither does it invariably “go along with the crowd” – whether it be a high school clique or company headquarters or the Marine corps. Differentiated autonomy maintains boundaries between itself and other individuals and groups, keeps its own counsel, and from that posture decides whether and how to relate to them. In the previous section, we have seen how many Christian philosophical theologians would insist that it is differentiated autonomy that will stand before the great judgment seat of Christ on the Last Day, differentiated autonomy that will be rewarded or punished for its self-determined choices for or against God.

3.3 The Limits of Autonomy

Without denying that differentiated autonomy represents an advanced stage, Christian philosophical theology has reason to deny that it represents the goal of human development. For personal creatures are meant to imitate God's stable orientation to the Good. But even virtue-enhanced differentiated autonomy is not a center that will hold. Further reorganization of

personal subjectivity is required, one that replaces the autonomous ego as manager of the personality with friendship with God as its functional core. John's Gospel argues this case by way of narrative illustration. Happily, we do not have to take a position on the historicity of the stories to find them conceptually clarifying.

3.3.1 *Jesus as Role-Model*

In John's Gospel, Jesus in his human nature is showcased as the paradigm "image and likeness" of God. Certainly, Jesus is portrayed there as a mature adult. So far as the psychological development of his human nature is concerned, Jesus has moved beyond identification with human authority figures, peer groups, or social institutions. John's Jesus teaches without footnotes (Jn 7:15). He is not narrowly tribal, but takes the initiative to convert Samaritans (Jn 4:7–9, 39–40), welcomes Greeks (Jn 12:19–23), and reassures Pilate that his kingdom is not of this world (Jn 18:36). What is important to John's Jesus is not social location or ethnicity, but how a person responds to ultimate values (Truth and Light). Certainly, John's Jesus does not hesitate to contradict the religious establishment who retaliate by plotting his execution. In relation to other human beings, he has achieved the competencies associated with Kegan's stage of differentiated autonomy.

Nevertheless, the functional core of Jesus's human personality is not the autonomous ego but his lived partnership with God the Father. In the language of John's Gospel, Jesus "abides" in the Father and the Father "abides" in him. They mutually "indwell" one another. Among other things, this means that Jesus's partnership with the Father is constant and voluntary. His frequent and intimate communication with the Father involves sharing points of view and harmonizing objectives. Jesus has entered into and embraced the Father's purposes. They are in complete agreement about his soteriological role. Everything Jesus says or does flows from as an expression of their ongoing interaction (Jn 5:19–23, 30; 6:38; 10:14–30; 12:27–28). Intimate trafficking with the Father wins Jesus's trust and acquaints him with Divine power to lay down life and to take it up again (Jn 5:24–29; 6:38–40; 10:17–18). Confident of this, Jesus can see the cross as his hour of glory, the culmination of his friendly faithfulness to the Father and the completion of the special work he had been given to do (Jn 12:27–28; 13:31–32). Experience of Divine favor and eternal life-preserving power is what stabilizes Jesus's orientation to the Good. Put otherwise, it is the perichoretic restructuring of his personality – the displacement of the autonomous ego with Divine friendship at his core – that is perfect freedom.

3.3.2 *The Process of Stage-Transitioning*

Other characters in John's Gospel are also presented as having in effect reached Kegan's state of differentiated autonomy. The disciples, the Pharisees and Sadducees, perhaps some of the people Jesus heals, are religiously serious people who have steeped themselves in the law and the prophets. They are religiously observant people who have habituated themselves to Godward practices. All of this is what makes them ripe for the next stage-transition. Jesus invites them to move into it (Jn 3:3). The Gospel narrative and discourses show what happens when mature virtue-enhanced autonomy opens itself up to, and what happens when it digs in against the reorganization of personal subjectivity around friendship with God.

3.3.2.1 *The Disciples* The disciples are ready and willing but – like adolescents reaching for adulthood – initially uncomprehending. John's Gospel introduces them as those who ask, who get to come and see where Jesus abides. At the core of his personality, Jesus makes

himself at home with the Father. This is the disciples' developmental goal (Jn 14:15–26; 15:26; 16:7, 12–15), but at the beginning they do not recognize it as such. At first, they confess their confidence in Jesus under superficial titles. Life together with Jesus – following where he leads, watching what he does, listening to what he says – catapults them into ever deeper recognition. Believing in Jesus is key, because Jesus is the divinely authorized manifestation of the Father (Jn 5:30–47; 7:16–29; 8:12–30; 12:44–50; 14:8–11). Because what Jesus says and does all flows out of his constant interaction with the Father, the disciples can get familiar with who the Father is by getting to know Jesus. The disciples' grasp of Divine intentions moves from the outside in, from the abstract to concretely embodied when they obey God's commandments (Jn 14:15–24). They get to know what Jesus and the Father are like as persons by joining in their projects, by “acting out” their intentions, by becoming more and more like them themselves!

Commandment-talk fits the pedagogy of transition, but it misdescribes the developmental goal. Commandment-keeping is *too oppositional* for friendship. When the disciples become friends of God and carry out Divine intentions or go along with God's program, they will be like John's Jesus: participating in projects regarding which they have given input (Jn 15:26; 16:23–24) and about which they have been consulted, taking up agreed roles and carrying out common objectives. Such “service” will be perfect freedom.

Talk of commandment-keeping misleads another way by suggesting a *division of labor*: that it is God who commands and the creature who follows Divine orders. The Gospels' lists of commandments ought to convince Christian philosophers: *there is not enough to even virtue-enhanced autonomous egos to obey them. Compliance can reliably happen only when personality is restructured into friendly collaboration with God!* Thus, the Synoptic Jesus makes clear: would-be disciples will have to move beyond *lex talionis* “do as you're done by” to the Golden Rule (Mt 7:12). On another occasion, Synoptic Jesus forwards the first and second great commandments: to love God with our whole selves and to love our neighbors as ourselves (Mt 22:37–40). Torah had already specified that reference to “neighbor” reaches beyond tribe and nation to include resident aliens (Lev 19:33–34). Torah makes clear: to love in this sense is to behave as though one has nothing to gain by denying others the necessities of life or access to life-giving resources and opportunities. The Good Samaritan midrash universalizes that injunction (Lk 10:35–37). Sermon-on-the-Mount Jesus goes further, commanding disciples to love their enemies and bless those who curse them: that is, to act as if they had nothing to gain by denying their arch-enemies (Mt 5:43–45). John's Jesus focusses on friends but ups the *ante*: for the disciples to love God is – among other things – for them to love one another as Jesus has loved them. It is to be persons prepared to lay down their lives for friends (Jn 15:12–14; cf. 10:15–18). What God expects of human beings, what Jesus demands of disciples goes beyond garden-variety moral virtue. Disciples are called to become saints and martyrs, whose whole manner of being and doing bears witness to who God is.

Experience shows that the vast majority of autonomous adults are unable to lay down life for friends, much less to love arch-enemies as themselves. Indeed, it is a rare few that show any inclination to do so. The narrative of John's Gospel bears witness to what is even more shocking: merely living at close quarters with Jesus on the outside, while attempting to learn the lessons he was teaching, was not enough to enable most of his disciples to remain loyal in the time of trial. Judas betrayed; Peter denied; the rest scattered in terror (Jn 18:2–27; 16:31–32). *John's Gospel is showcasing how – even with the religiously serious and well intentioned who open themselves to stage-transition – virtue-enhanced differentiated autonomy is not a center that will hold.* Over the weekend of Jesus's crucifixion, the disciples' integrity went to smash in a developmental crisis of horrendous proportions. Happily, in John's Gospel, the risen and

ascended Jesus reappears on Easter Sunday evening to resurrect them (Jn 20:19–23). With their old subjectivities dismantled, their personalities can be reorganized to put friendship with God at their core (Jn 14:15–28; 17:23, 26). Lived partnership with the One who has power to lay down life and power to take it up again is required to stabilize their orientation to the Good and so to usher them into perfect freedom.

3.3.2.2 *The Resisters* In John's Gospel, the Pharisees and Sadducees are also virtue-enhanced autonomous adults ripe for stage-transition. But most of them "dig in" against Jesus's invitation and cling to their autonomy. It is a commonplace of developmental psychology: entrenched refusal to advance risks regression. In John's Gospel, the religious leaders betray their deepest loyalties by plotting the crucifixion of the Messiah whose way they had meant to prepare. The religious zeal that seems so virtuous gets disastrously misdirected when they refuse to let it be recontextualized in restructured personality.

3.4 *Metaphysical Contrasts*

Christian philosophers trying to solve the problem of evil by identifying some aspect of created agency that not even an omnipotent God can control take their metaphysical stand with incompatibilism. For humans, being made in God's image consists in being rational creatures who act through reason and incompatibilist free will. By contrast, identifying the image of God with personal subjectivity emphasizes its developmental and interactive nature. Even developmental psychologists who posit autonomy as the goal of human development would not understand it in terms of source-incompatibilism. To the extent that they think about the metaphysical question at all, most take some sort of non-reductive physicalism for granted. That is, they would assume that human agency is at bottom constituted by microphysical particles whose interactions are governed by the laws of physics. Accordingly, they would reject source-incompatibilism and instead conceive of personal agency in terms of the state of macro-level causal integration and dynamic functioning. Put otherwise, they would see human persons as the focal source, not the ultimate source of their actions. What developmental psychologists chart is the successive ways in which personal subjectivity gets focused. What they borrow from Kant is not his metaphysics of freedom, but his portrait of the autonomous agent's functional capacities for practical rationality – the capacity to formulate and commit oneself to principles as rationally binding, the capacity to set goals and calculate apt means in accord with principles that dictate respect for all humanity. Pereboom contends that these constitute the core of Kantian autonomy, and that they are compatible with physicalism.³⁴ Recognizing perichoresis as a further developmental stage does nothing to shift the metaphysical presumptions. It only reidentifies the focus from which truly free acts proceed.

3.5 *Regression or Advance?*

Metaphysics aside, psychologists and moralists might deplore the idea of perichoretic personality as developmentally regressive. Kegan might protest: after all that work to achieve adult differentiation, are we called to re-embed in the matrix of the Trinity to let our words and deeds be the resultant of communal interaction? The "size-gap" between infinite Godhead and finite humanity guarantees that any friendship with God will be overwhelmingly lop-sided. Why isn't this just backsliding into a construction of the self that is dominated by authority figures or that goes along with the crowd? Christian philosophical theology does not share Kegan's Kantian bias against embeddedness, however. Christian philosophical

theology affirms Godhead as a Trinity of persons, a personal matrix of identity-conferring relationships. The Western theological tradition holds that the persons of the Trinity are metaphysically distinct, *and* that it is metaphysically impossible either that they should exist separately or that they should fail to agree. Already in the eleventh century, Anselm declares: one of God's goals in creation is that some creatures should be as godlike as possible. Embedded in the Trinitarian matrix, making friendship with God the functional core of one's subjectivity, *mutatis mutandis* seeing as the Trinity see and loving as the Trinity love, is as godlike as it would be possible for a personal creature to be.

Making friendship with God one's functional core does not mean giving up the adult competencies that go with differentiated autonomy. Precisely because of the size-gap, partnership with God does not put Divine and human agencies into competition for control, as if they were mere peers. Rather God the all-wise Creator is an agency-enabler who has no interest in stunting human growth. God wants human beings to grow up to full stature, to work their way through to differentiated autonomy in relation to other human beings. Friendship with God at the core does not step back from that. It is that fully differentiated adult self that is ripe for stage-transition. Partnering with power to lay down life and take it up again does not rob us of our adulthood, but stabilizes our orientation to Goodness by depriving this-worldly threats of their power to intimidate. Contemporary saints and heroes join John's Jesus to convince us. Mahatma Gandhi, Martin Luther King, Desmond Tutu, Mother Teresa, and the Dalai Lama are certifiably mature adults who have been nevertheless been converted until who they are, what they say and do, become remarkable expressions of Divine love.

The Editors wish to express their deep gratitude to Professor Robert Merrihew Adams for assistance and advice in finalizing this chapter authored by Professor Marilyn McCord Adams.

Notes

- 1 Anselm, *Monologion*, cc.4 & 31; Schmidt I.16–18, 47, 49. Aquinas, *Summa Theologica* I, q.15, a.1 ad 3um & a.2 c. Bonaventure, *In Quattuor Libros Sententiarum* I, d.35, qq. 1 & 3; Quaracchi I.600–601, 608–609. Henry of Ghent, *Quodlibeta* VIII, q.1 (Venice 1613), II, fol. 1v; *Quodlibeta* IX, q.2; Leuven XIII.26–33.
- 2 These claims underlie cosmological reasoning. See Anselm, *Monologion*, cc.1–4; Schmidt I.13–18. See also Scotus, *Ordinatio* I, d.2, p.1, qq.1–2, nn.39–156; Vat II.148–221.
- 3 *Peter Abelard's Ethics*, ed. with introduction and translation by D.E. Luscombe (Oxford: Clarendon Press, 1971).
- 4 Augustine, *De Libero Arbitrio*; *Corpus Scriptorum Ecclesiasticorum Latinorum* 74 (Vindobonae: Hoelder-Pichler Tempsky, 1956) I.86.26; I.101.30–31; III.33.97–98. Scotus, *Quaestiones super libros Metaphysicorum Aristotelis* IX, q.15, a.2, nn.20–34; OPH IV.680–684. Ockham, *Scriptum in I Sententiarum*, d.1, q.6; OTh I.501–502; *Quaestiones in III Sententiarum*, q.6; OTh VI.170–176; *Quaestiones in III Sententiarum*, q.11; OTh VI.375; *Quaestiones in IV Sententiarum*, q.16; OTh VII.359; *De Connexione Virtutum*, q.7, a.4; OTh VIII.381; *Expositio in libros Physicorum Aristotelis*, II, c.8, sec.1; OPh IV.319–320.
- 5 Ockham, *Ordinatio* I, d.1, q.6; OTh I.490–496, 501–502; *Ordinatio* I, d.2, q.1; OTh II.10–11; *Ordinatio* I, d.10, q.2; OTh III.340–343.
- 6 There he declares that the action by which God creates is expressing his word (an act of thinking), that the word by which God expresses creation and the word by which God expresses Godself are identical, that God eternally expresses Godself in such a way that there is no power in the universe to prevent it. There is no mention of Divine will in this context. See *Monologion*, cc.12, 29–30, 32–33; Schmidt I.26, 47–48, 51–53.

- 7 Peter Abelard, *Theologia Christiana* V, secs. 31–32; *Corpus Christianorum Continuatio Mediaevalis* XII (Brepol: Turnholt, 1969), 359. Aquinas reviews such arguments in *Summa Theologica* I, q.25, a.5, args. 1–3, and rejects them.
- 8 Aquinas, *De Potentia*, q.3, a.5 ad 14; *Summa Contra Gentiles* II, c.31; *Summa Theologica* I, q.19, a.3 c & ad 2um.
- 9 Scotus, *Lectura* I, d.39, qq.1–5, nn.41–90; Vat XII.492–509; *Ordinatio* I, d.8, p.2, q.u, nn.281–291; Vat IV.313–321; *Ordinatio* II, d.1, q.2, n.70; Vat VII.38–39.
- 10 Ockham, *Quodlibeta* II, q.9; OTh IX.154–155.
- 11 Anselm, *De libertate arbitrii*, cc.1–2; Schmidt I.207–210. *De Casu Diaboli*, cc.2–15, 21–25; Schmidt I.251–259, 266–273.
- 12 For example, see Aquinas, *Summa Theologica* II-1, qq.6, 8, 9, 13, 18, 76–78. The precise classification of Aquinas's action theory is vexed. But many passages require no more than focal sourcing.
- 13 Augustine, *De Libero Arbitrio* I.109.32–33; *De Civitate Dei* V, c.10, n.1 (PL 41.152; CSEL 40–1.229). Ockham wrestles to reconcile authorities who say that rational creatures with confirmed orientation to the good are freer, with his own conviction that they retain a liberty of indifference with respect to many choices (*Ordinatio* I, d.10, q.2; OTh III.341–344).
- 14 Anselm, *De Libertate Arbitrii*, c.1; Schmidt I.212; *De Concordia Praescientiae et Praedestinationis et Gratiae Dei cum Libero Arbitrio* I.6; Schmidt II.257.
- 15 Ockham, *Quaest. in IV Sent.*, q.16; OTh VII.351; *Quodlibeta* II, q.9; OTh IX.154–155.
- 16 Bonaventure gives his version of this story in *Breviloquium*, Part III, cc.1–11; Quaracchi VI.231–241. See also Marilyn McCord Adams, “Theodicy without Blame,” *Philosophical Topics* XVI (1988), 215–245; esp. 218–222.
- 17 Alvin Plantinga, *The Nature of Necessity* (Oxford: Clarendon Press, 1974), ch.9, 164–193.
- 18 J.L. Mackie, “Evil and Omnipotence,” *Mind* 64 (1955), 200–212.
- 19 For a recent nuanced development of this line, see Kevin Timpe, *Free Will in Philosophical Theology* (New York and London: Bloomsbury Publishing, 2014).
- 20 For an extended analysis of Scotus' attempts, particularly in *Reportatio* IA, d.39, see Allan B. Wolter, “Scotus's Paris Lectures on God's Foreknowledge of Future Events” in *The Philosophical Theology of John Duns Scotus*, ed. by Marilyn McCord Adams (St. Bonaventure, NY: Franciscan Institute Publications, 2015), 381–437.
- 21 See Ockham, *Tractatus de Predestinatione et de Praescientia Dei respectu Futurorum Contingentium*, OPh II.509–539. See also Marilyn McCord Adams, *William Ockham* (Notre Dame, IN: University of Notre Dame Press, 1987), ch.27, 1115–1150.
- 22 For explanations of Molinism, see Plantinga, *The Nature of Necessity*, ch.9, and Alfred Freddoso's introduction to his translation *Luis de Molina: On Divine Foreknowledge: Part IV of The Concordia* (Ithaca and London: Cornell University Press, 1988), 1–81. For critiques, see Robert Merrihew Adams, “Middle Knowledge and the Problem of Evil,” *American Philosophical Quarterly* 14.2 (1977), 109–117; and William Hasker, *God, Time, and Knowledge* (Ithaca and London: Cornell University Press, 1998).
- 23 For explanations and defenses of Open Theism, see William Hasker, *God, Time, and Knowledge*; Clark H. Pinnock, *Most Moved Mover: A Theology of God's Openness* (Grand Rapids, MI: Baker Books, 2001); and *God in an Open Universe: Science, Metaphysics, and Open Theism*, ed. by Thomas Jay Ord, William Hasker, and Dean Zimmerman (Eugene, Oregon: Pickwick Publications, 2011).
- 24 Scotus, *Ordinatio* I, d.17, p.1 q.2, nn.31–50, 87–91, 151; Vat V.159, 181–184, 211. Ockham, *Quaestiones in II Sententiarum*, q.15; OTh V.340; *Quaestiones in III Sententiarum*, q.11; OTh VI.357–358.
- 25 Ockham, *Ordinatio* I, d.1, q.2; OTh I, 397, 399; *Quaestiones in II Sententiarum*, q.15; OTh V.341; *Quaestiones in III Sententiarum*, q.11; OTh VI.357–358.
- 26 David Lewis, “Evil for Freedom's Sake,” *Philosophical Papers* 22 (1993), 149–172.
- 27 See Marilyn McCord Adams, *Horrendous Evils and the Goodness of God* (Ithaca and London: Cornell University Press, 1999), ch.3, 32–39.
- 28 This term is a familiar one in Trinitarian studies, and refers to the mutual indwelling of Jesus and the Father, as discussed in Section 3.3.1. For readers unfamiliar with the term, see the entry “Per-

- ichoreisis,” Wikipedia article <https://en.wikipedia.org/wiki/Perichoreisis> (accessed on 9th August 2020). [Editor’s note inserted with permission and advice of Professor Robert M. Adams, 2020.]
- 29 Within analytic circles, Thomas Nagel reasserted the importance of the subjective point of view in his “What is it like to be a bat?” *Philosophical Review* 83, 435–450. For a more recent exploration of personal subjectivity, see L.A. Paul, *Transformative Experience* (Oxford: Oxford University Press, 2014).
 - 30 See Erik H. Erikson, “The Eight Stages of Man” in *Childhood and Society* (New York and London: W.W. Norton & Company, 1950), ch.7, 247–274. Erikson correlates each stage with a dilemma: Trust vs Mistrust, Autonomy vs Shame and Doubt, Initiative vs Guilt, Industry vs Inferiority, Identity vs Role Confusion, Intimacy vs Isolation, and Generativity vs Stagnation. Successively striking balances in favor of the first pole builds towards healthy adulthood.
 - 31 James E. Loder, *The Transforming Moment: Understanding Convictional Experiences* (San Francisco: Harper & Row, 1981).
 - 32 Lawrence Kohlberg, *The Philosophy of Moral Development: Essays on Moral Development* (San Francisco: Harper and Row, 1981 & 1984), vols.1–2.
 - 33 Robert Kegan, *The Evolving Self: Problem and Process in Human Development* (Cambridge, MA and London: Harvard University Press, 1982). See also James W. Fowler, who draws on work by Piaget, Kohlberg, and others to chart the stages of faith (*Stages of Faith: the Psychology of Human Development and the Quest for Meaning* [New York: Harper Collins, 1981]).
 - 34 Derk Pereboom, *Living without Free Will* (Cambridge, UK: Cambridge University Press, 2001), ch.5, 151–153.

Part II

Compatibility Problems

The Consequence Argument and the *Mind* Argument

JOSEPH CAMPBELL AND KENJI LOTA

1 Introduction

Free will is an enigma. There are reasons to suppose that free will is incompatible with determinism, and other reasons to suppose that we lack free will if determinism is false. Combining the arguments, we get a simple constructive dilemma.

- (I) If determinism is true, then no one has free will.
- (II) If determinism is false, then no one has free will.¹
- (III) Determinism is either true or false.
- (IV) Ergo, no one has free will.

In this essay, we provide formal versions of arguments for premises (I) and (II), i.e., the *consequence argument* (CA) and the *Mind argument* (MA), respectively. Our study and comparison of the arguments suggests they have the same determinism-independent core. This is an odd result, and we briefly discuss how to interpret the findings in the final section.

Some terminology and clarifications are needed before we begin. *Free willism* is the view that some persons have free will, where a person has *free will* provided some actions or events are up to her. *Free will skepticism* is the denial of free willism, *compatibilism* holds that free willism is consistent with determinism, *incompatibilism* is the denial of compatibilism, and *impossibilism* is the view that free willism is impossible (van Inwagen 1983; Campbell 2011; Vihvelin 2013).

Broadly speaking, there are two theories of free will. The *source view* maintains up-to-usness, or sourcehood, properly described is enough to characterize free will, but the *leeway view* adds the additional requirement that the agent be *able to do otherwise* (cf. Pereboom 2001). For Peter van Inwagen, persons with free will have “the *power or ability* to act otherwise than they do” (1975, p. 188). Free will requires alternative possibilities, the *specific ability* to perform either of two actions at a particular time (Campbell 2013). Leeway free will – which requires alternative possibilities – seems to clash with determinism – understood as a kind of necessity. Looking at it in this way, CA is part of a history of modal and philosophical logic dating back to antiquity (Kneale and Kneale 1962, Chapter III).

Note that on this understanding, leeway free will is a *dual-ability* involving both the ability to act and the ability to act otherwise. Perhaps the concept of *dual-ability* is incoherent for the ability to do otherwise is incompatible with determinism and the source component is incompatible with indeterminism.² Indeed that is one interpretation of our results. Similarly, source compatibilists might take this essay as evidence against the leeway view. These and other interpretations are discussed at the end of the chapter (§6). The source view and free will skepticism garner more support than the leeway view among contemporary philosophers, but we should all be interested in the logic of leeway free will if only to better understand its failings.

Van Inwagen (1983) was the first to develop and compare formal versions of CA and MA, and his work is the main inspiration for this chapter. We begin with CA.

If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us. (1983, p. 16)

The *remote past* is the time prior to the existence of human beings (Finch and Warfield 1998). According to the proponent of CA, propositions about the past and the laws are not up to us for we cannot render them false. Determinism extends this inability to all true propositions.

Indeterminism is far more complicated since it is merely the denial of determinism and may be manifested in several different ways. We are concerned with the potential incompatibility of free will and indeterminism, how indeterminism might pose problems for free will. For this reason, we restrict the thesis of indeterminism throughout. Dana Nelkin clarifies: “if indeterminism is to be relevant to the formation of free actions, it *must* be that actions are caused by, but not determined by, agents’ mental states and deliberations” (2001, p. 109, fn. 4). This is phrased in terms of event-causal theories (Kane 1996; Balaguer Ch. 19), and for simplicity we generally follow this practice, but the point is to restrict indeterminism so as to rule out worlds with spatiotemporally distant indeterministic events in an otherwise determined system (van Inwagen 1983, p. 126). The question about the relevance of determinism or indeterminism to free will is a question about the pertinent causal relations between the agent and her action, and whether they are determined. The indeterminism with which we are concerned is manifested locally, and events leading to actions are assumed to follow in accordance with probabilistic laws. This carries with it a presumption of *nomological naturalism* – that the future is a nomological consequence of the past, that is, future actions are “caused by, but not determined by, agents’ mental states and deliberations.” Nomological naturalism helps to highlight the comparison between CA and MA.³

Van Inwagen offers an array of versions of CA and MA. In Chapter III of *An Essay on Free Will* (1983), he advances three formal versions of CA. He writes: “any specific and detailed objection to one of the arguments can be fairly easily translated into specific and detailed objections to the others” (1983, p. 57). In contrast, Chapter IV is devoted to arguments for compatibilism. Among them is at least one strand of MA squarely in support of premise (II). Versions of MA are called “strands” since they differ in important ways but share the same general goal. Specifically, they are all supportive of compatibilism and critical of *libertarianism*: the view that indeterminism is true because free willism and incompatibilism are true.

These remarks suggest that van Inwagen regards the different versions of CA as *equable*, that is, if one of the arguments is sound, then the other versions must also be sound.⁴ This can be weakened to apply more generally and in different contexts, but it is suitable for the purposes of our essay. Van Inwagen comes close to suggesting that the third version of CA (CA-3) is equable

to the third strand of MA (MA-3). At least he cannot spot any reason for accepting one argument and rejecting the other (1983, pp. 145–152). For libertarians like van Inwagen who accept free willism yet are incompatibilists often because of CA, the equability of CA-3 and MA-3 is problematic. Forced to choose between “the inconceivable and the puzzling” (1983, p. 150), van Inwagen chooses the puzzling. He is confident of the soundness of CA-3, and he is more confident there is a mistake in MA-3 than he is that no persons have free will (2004, pp. 224–225). Van Inwagen’s *mysterianism* hinges on there being a substantive difference between the arguments.

Using van Inwagen’s definitions and methodology, we construct formal versions of CA (§§2–3) and MA (§§4–5) that rely on the same assumptions and inference rules. We conclude that two formal versions – CA-3 and MA-3 – are equable because they share the same basis: the *core of CA* (or *core of CA/MA*). This allows for a determinism-independent argument for free will skepticism (§5). We finish with a brief summation and some plausible interpretations of our results (§6). For now, it is best to understand our thesis as *if CA-3 is sound, then so is MA-3*. Our goal is to expose the tension of adopting conflicting attitudes about the relative merits of CA and MA. We ignore some of the better criticisms of CA-3 for this reason (see especially Ismael 2016).

2 Support of Premise (I): CA

In this section and the next we introduce and develop formal versions of CA in support of premise (I).⁵ Carl Ginet (1966) and Wilfred Sellars (1966) deserve equal credit for the first formal discussions of CA, but informal versions of the argument have a history dating back to antiquity (Long and Sedley 1987, pp. 186–187). Van Inwagen (1983: Ch. III) offers three formal versions of CA. He attempts to establish that if determinism is true, then – given other reasonable assumptions – no one has a choice about anything, so no one has free will.

There are two standard ways of defining determinism. *Semantic determinism* is the conjunction of two theses.

- For every instant of time, there is a proposition that expresses the state of the world at that instant;
- If p and q are any propositions that express the state of the world at some instants, then the conjunction of p with the laws of nature entails q (1983, p. 65).

As defined, determinism expresses a symmetrical relation between propositions. Elsewhere, van Inwagen defines *metaphysical determinism*, which is less formal and asymmetrical: “there is at any instant exactly one physically possible future” (1983, p. 3). We follow van Inwagen and take these to be comparable definitions, though not everyone agrees (Mickelson Ch. 4).

How does the semantic story fit with the metaphysics? David Lewis writes: “Let us say that an event would *falsify* a proposition iff, necessarily, if that event occurs then that proposition is false” (1981, p. 119; our emphasis). Similarly, John Perry (2004) distinguishes between propositions being *made true* and being *settled*. Suppose a man, Elwood, is causally determined to not eat a cookie at some future time, t .

Perry writes:

I will say that although the proposition that Elwood will not eat the cookie at t has not yet been *made true*, its truth value has been *settled*. ... To be settled at t is to follow from some set of propositions, each of which is either established or made true by time t . (pp. 236, 250)

Actions are events, too. By performing actions, we make some propositions true and falsify others. The problem is that some propositions are settled by previous events and their consequences. The compatibility problem comes down to whether we can falsify any true propositions given determinism, for in that case all the true propositions appear to be settled.

Versions of CA are distinguishable in several ways. First, there are *transfer versions* and *extension versions* (Campbell 2017). The former arguments rely on transfer principles like principle (β) (introduced below) whereas the latter rely on principles that limit relevant possibilities to physically possible extensions of the past, such as:

Extension: Person S is able to do action X at time t only if that S does X at t is consistent with the conjunction of propositions about the past (relative to t) together with the laws of nature. (cf. Fischer 1994, p. 88; Haji 2009, p. 53)

Supposing extension were true, it is easy to see that determinism and free will conflict, for given extension and determinism it follows that no one is able to do otherwise, so no one has (leeway) free will. CA-2 – van Inwagen’s second argument (1974, 1983: §§3.6–3.9) – is an important extension versions of CA and there are others by Ginet (1990), John Martin Fischer (1994: Ch. 5), and Ishtiyaque Haji (2009: Ch. 2).

There are also *individual* and *general versions* of CA (Campbell 2017). The former attempt to show that a particular person in a particular situation cannot do otherwise given determinism. Provided that the person and situation are arbitrary, we may generalize and conclude that no one can do otherwise given determinism. This is how we interpret CA-1 – van Inwagen’s first formal version of CA (1975, 1983: §§3.3–3.5). Van Inwagen provides an example of a judge, J , who fails to raise his hand at some time T . CA-1 attempts to establish that J could not have raised his hand at T given determinism (1983, p. 68ff.). If the person and the situation are indeed arbitrary, and there are no advantages bestowed or questionable assumptions at play, then incompatibilism follows.

There is some dispute about the details of CA-1. We disagree with Ted Warfield (2000) who charges that there is a “modal fallacy” in versions of CA. His contention is that the necessity of the premises, not merely their truth, must be established in arguments for incompatibilism, since the conclusion is held to be true in all possible worlds (2000, p. 170). Van Inwagen (2000, p. 18, footnote 2) shows a sympathy with Warfield:

‘ P_0 ’ is a contingent truth, and ‘ NP_0 ’, which has ‘ P_0 ’ as a conjunct, is therefore a contingent truth.
‘ L ’ is probably a contingent truth, and ‘ NL ’ is therefore probably a contingent truth.

In both CA-1 and CA-3, “ P_0 ” denotes a proposition that expresses the entire state of the world at a time t_0 in the remote past, and “ L ” denotes the conjunction of the laws of nature. “ N ” is the no-choice operator, read roughly, “no one has a choice about whether ...” P_0 is contingent because it says something specific about the remote past of the actual world, specifying a proposition that might be false in other possible worlds. Even if other possible worlds have a remote past, P_0 might not express a true proposition in that world. To formulate a more “careful statement,” van Inwagen writes:

... an argument identical in appearance with the argument in the text can be constructed in any possible world and [the] premises ... of any of these arguments will be true in the possible world in which it is constructed if ‘ P_0 ’ expresses a proposition that describes the state of the world (= ‘universe’) in that possible world before there were any human beings, and ‘ L ’ expresses the proposition that is the conjunction of all propositions that are laws of nature in that possible world. (2000, p. 18, footnote 2)

We concede the truth of these clarifications, but the point is overstated.

In *universal generalization*, showing something to be true for an arbitrary person, shows that it must be true for all persons. Similarly, in a *modal generalization* showing that something is true at an arbitrary world, shows that it must be true in all worlds, and thus is necessarily true. There are multiple levels of generalization in CA-1 and CA-3. What is important is whether there are assumptions that allow for arbitrary advantages applicable to some but not all people, or applicable in some but not all possible worlds. All assumptions must be generalizable. We return to this point below.

CA-3 is more general than CA-1, though generality is a matter of degree. What matters is that individual transfer versions incorporate individual transfer principles, that is, principles that are indexed to individual persons perhaps to individual persons performing actions at particular times (Ginet 1966; Fischer 1994). More general transfer versions like CA-3 use more general transfer principles. These four types of argument – transfer vs. extension, individual vs. general – can be mixed and matched. For instance, CA-2 is a general extension version of CA framed in first-order logic.

CA-3 (1983: §§3.10–3.11, 1989, 1994, 2000, 2015) is our favored formal version. It incorporates a *no-choice operator*: “ Np ” stands for “ p and no one has or ever had a choice about whether p .” *Having a choice* is taken to mean having the ability to *render a true proposition false*. In preliminary versions of CA-3, we adopt the *strong sense* of “render” (Lewis 1981), *to act so as to ensure that a true proposition is false* even if the proposition is or will be true (see van Inwagen 1983, pp. 67–68, and fn. 31, pp. 233–234; Finch and Warfield 1998). We discuss the weaker sense below. Following van Inwagen (2015), a proposition is *inevitable* if and only if it is true and no one is or ever was able to (strongly) falsify it. That every true proposition is and always has been inevitable is the conclusion of the *inevitability version* of CA-3.

Let “ $\Box p$ ” stand for “ p is broadly logically necessary” and “ $\supset$ ” designate the relation of material implication. “ P_o ” and “ L ” are defined as above, respectively, a proposition that expresses the entire state of the world at a time t_o in the remote past, and the conjunction of the laws of nature. “ F ” denotes an arbitrary, true proposition about the future. The argument uses two inference rules:

- (α) From $\Box p$ deduce Np ;
- (β) From Np , $N(p \supset q)$ deduce Nq .

Here is a short version of CA-3 (cf. Nelkin 2001, p. 108; Widerker 1987).

- (1) $\Box((P_o \& L) \supset F)$ assumption (df. ‘determinism’)
- (2) $N((P_o \& L) \supset F)$ from (1) by (α)
- (3) $N(P_o \& L)$ assumption (necessity of the past & the laws)
- (4) $\therefore N(F)$ from (2), (3) by (β)

Since F is an arbitrary truth, we may generalize the conclusion: No one is able to (strongly) falsify any true proposition. Determinism entails that every proposition is inevitable.

In van Inwagen’s version of CA-3 (1983, pp. 94–95) there are more steps because he separates premise (3) into two different claims, one about the necessity of the past and another about the necessity of the laws of nature. We phrase the argument in terms of *the necessity of the past and the laws* for reasons that will be clear later in the essay. Determinism must be accepted for the purposes of conditional proof, so if the inference rules are sound, the burden of the argument shifts to premise (3). Of course, one may reject the premise for at least two different reasons: one may reject the necessity of the past, or one may reject the necessity of the laws.

Debates about the necessity of the past are complex and often confused. In an important sense, premise (3) is not really about the necessity of the past, for the operative claim is about the remote past, the past prior to the existence of human beings. Aidan can do nothing this evening to ensure that she did not wear converse sneakers this afternoon (given the time has passed, and she did wear the sneakers). However, earlier this morning she might have been able to ensure that she did not wear converse sneakers this afternoon. What makes the remote past salient is not that we can't do anything about it now, but that no one could have done anything about it then, when the events of the remote past were present. The remote past is prior to the existence of human beings, so no one was available to do anything.

However, that there is a remote past is a contingent feature of the universe. Nor is P_o contingent merely in the sense that L or F are contingent. The laws of nature might be different in different possible worlds, so in some worlds they won't be expressed by " L ." But there are laws of nature in all possible worlds. P_o is not merely contingent but it is a non-arbitrary assumption, one that does not hold in all possible worlds, or at least in all worlds that determinism is true. Some possible worlds do not contain propositions of that type, because they lack a remote past (Campbell 2011, p. 51). Warfield (2000) is correct in noting that formal versions of CA contain spurious contingencies, but he gets the nature of the contingencies wrong.

The problem with CA-1 and CA-3 is that some of the assumptions are not arbitrary for they apply to only a subset of possible worlds. P_o only denotes a proposition in possible worlds with a remote past, but not all worlds have a remote past. CA-3 fails to establish incompatibilism, i.e., there are no possible worlds where both free willism and determinism are true, since worlds without a remote past are excluded from consideration (Campbell 2007). Andrew Bailey (2012) calls this the *no-past objection*, and neither Warfield nor van Inwagen confront it. Bailey (2012) and Jiji Zhang (2013) extend the results of the no-past objection to other arguments for incompatibilism, such as the manipulation argument. The objection applies to any argument that depends on the existence of a pre-human history.⁶

In response to the no-past objection, Carolina Sartorio claims that what matters is that our "causal history is one that traces back to factors beyond our control" (Sartorio 2016, p. 265). Following Sartorio, since the past and the laws are beyond our causal control (Pereboom 2014, p. 1), both are necessary in this no-choice sense. Principle (β) transfers this lack of choice and control onto other true propositions. Given determinism our lack of choice is settled by propositions about the remote past and their nomological relationship to the future, all of which is beyond our control.

Some important points are being side-stepped in the above reply. The central conclusion of the no-past objection is that neither CA-1 nor CA-3 are demonstrations of incompatibilism. Warfield was correct about that. At most, they prove conditional claims: If there is a remote past and determinism is true, then no one has free will. Also, if the past is necessary, why should its remoteness matter? (Campbell 2008, 2010) Most importantly, claims about the remote past are contingent whereas the conclusion purports to convey a kind of necessity that applies to all propositions: no-choice or inevitability, as we currently construe it. We should be suspicious of CA-1 and CA-3 for the same reason David Hume (1947, 1978) was suspicious of demonstrations of existential claims. There is a failure of closure. What we have is not a demonstration of incompatibilism but a *demon-stration*: a necessary conclusion based on a contingent, arbitrary assumption. Proponents of incompatibilism should strive for a better argument.

Our interests are in the merits of CA-3 relative to MA, so we accept Sartorio's response. With the response comes a picture of how formal versions of CA work. Given determinism, our actions trace back to the past and the laws – factors beyond our control. In CA-2, we

convey this worry with the use of principles like extension, but in CA-1 and CA-3 we break it down further. There is a need for *grounding principles*, which fix our lack of choice, and *transfer principles* like (β) that transmit the property of no-choice onto other true propositions (Campbell 2011, p. 51). Following Sartorio's lead, the grounding principles are about true propositions that are causally beyond our control such as propositions about the past and the laws.

We discussed one reason to deny premise (3) but there is a second reason, since (3) is about the past and the laws. The usual assumption in debates about free will is *realism* about the laws of nature, the view that the laws of nature have ontological priority over events. Events are governed by laws, and events occur *because* of the laws. According to *Humeanism*, this relationship is reversed. On one popular view, the laws of nature are empirically fruitful generalizations of past events, some of which are human actions (Hume 1975; Lewis 1986). Realists suppose the laws dictate the events of the world, but Humeans hold that laws of nature are grounded by events. Since some of the events are actions that are presumably up to us, one may argue that the laws are up to us, too.

This is one way of telling the Humean story, but not the only way (cf. Beebe and Mele 2002). However, there are features of laws of nature that foster a popular and contrary attitude in the context of debates about CA, which makes it difficult for Humeanism to gain traction. Consider van Inwagen's example of *intersidereal travel*, that is, travel between constellations. Intersidereal travel may be hindered in either of two ways: due to technology or due to the laws of nature. However, the speed of light marks an upper limit on space travel. There are no violations of laws of nature. No matter what advancements are made in space travel, intersidereal travel "will always be a matter of years or centuries." Van Inwagen concludes: "No technological advance could ever change this unfortunate fact, for it is a consequence of the laws of nature" (1983, p. 7).

In other words, though technology often provides us with the opportunity to falsify some generalizations (e.g., no person can fly), it will never allow us to (strongly) falsify a law of nature. This is clear given the strong sense of "render," for no one is able to act so as to *ensure* that a law of nature is false. Keith Lehrer (1980) and Lewis (1981) suggest that there is a weaker sense of "render" whereby we are able to falsify propositions about the past or the laws of nature, yet even on that story we are not going to perform actions that (strongly) falsify laws of nature, nor that (strongly) falsify material conditional instantiations of laws of nature. Lewis is very clear about this.

Had I raised my hand, a law would have been broken beforehand. The course of events would have diverged from the actual course of events a little while before I raised my hand, and at the point of divergence there would have been a law-breaking event—a divergence miracle, as I have called it. But this divergence miracle would not have been caused by my raising my hand. (1981, p. 117)

However the compatibilist is to escape the fine mess CA-3 has gotten her into, (strongly) falsifying laws of nature is not among those options.⁷

The suggestion that intersidereal travel is possible because one day we will violate laws of nature, or the consequences of any laws of nature, is regarded as a non-starter. Consequently, that we won't falsify (in the strong sense) a law of nature is a working assumption in discussions of CA. We grant the abstract possibility of Humeanism but it is practically useless in debates about the soundness of the inevitability version of CA-3. Perhaps we can (weakly) falsify laws of nature – "I am able to do something such that, if I did it, a law would be broken" (Lewis 1981, p. 115). But the laws are not broken in the actual world, nor will they be broken in the

future. The effects of Humeanism are *unrealizable*, that is, unavailable once we fill in the settled facts, because (strong) violations of law are not allowed. This has been built into the dialectic of debates about the soundness of versions of CA. Even Humean laws are never violated. We fail to realize that such an attitude – in a context where the focus is on the specific ability to do otherwise at a particular moment – smothers the Humean spark before it begins to ignite.

Note that our lack of control over these generalizations, given the story above, is independent of the determinism/indeterminism debate, and independent of the realist/Humean debate. It is because in the context of debate about the soundness of CA-3, we regard the laws of nature and the conditionals subsumed under them as settled facts. The debates concern the modal status and grounding of those facts, but the settled nature of their truth value is not up for debate. Not only does our inability to (strongly) falsify laws of nature – or to falsify any generalizations established by laws of nature – deflect Humean criticisms of premise (3), but it points to a way of gaining support for proposition (2) independent of premise (1). Recall that even if determinism is false, we retain the assumption of nomological naturalism – e.g., that events about agents are nomologically related to events about the future and that these relations are patterned by the lawlike connections between them. We add to this that we lack the ability to (strongly) falsify some of the true claims about these nomological relations, as well as the material conditionals entailed by them.

3 The Core of CA-3

One reply to the inevitability version of CA-3 is to reject premise (3) and argue for Humeanism. The Humean reply is not effective in this context, for reasons noted. Moreover, the above discussion raises a new question: What is the purpose of premise (1)? The tight relationship between the past and the future is the result of nomological relations not entailment relations. The material conditional in proposition (2) – $((P_o \& L) \supset F)$ – is grounded by the laws of nature for you cannot falsify the material conditional without falsifying one of the laws. But we can't (strongly) falsify laws of nature. Our inability to falsify propositions is more glaring in cases of entailment, but in this case entailment is intended to capture the underlying nomological relations between events. We use entailment only because causation is "a morass" (van Inwagen 1983, p. 65).

Note that the above considerations in support of proposition (2) logically separate it from principle (α). This allows for an even shorter version of CA-3.

(2) $N((P \& L) \supset F)$	assumption (necessity of nomological relations)
(3) $N(P \& L)$	assumption (necessity of the past & the laws)
(4) $\therefore N(F)$	from (2), (3) by (β)

The argument uses just one inference rule: principle (β).

We've made progress but before continuing we must confront a counterexample to principle (β). Tom McKay and David Johnson (1996) offer a counterexample to a related rule:

Agglomeration: From Np , Nq deduce $N(p \& q)$.

Van Inwagen sets up the counterexample.

Consider a coin that is never tossed. Suppose that I have a choice about whether the coin is tossed, but ... [not] about how the coin would fall if tossed. (van Inwagen 2015, p. 18)

Peter does not flip a coin. Since the coin is not flipped, it is inevitable *that the coin does not land heads* and it is inevitable *that the coin does not land tails*, but it is not inevitable *both that the coin does not land heads & that the coin does not land tails*. Peter could have flipped the coin, making the conjunction false. Agglomeration is derivable from principles (α) and (β), and since the former is assumed to be beyond reproach, principle (β) is rejected.

The example can be manipulated to provide a direct counterexample to (β) (van Inwagen 2015, p. 18). The success of the counterexamples hinges on indeterminacy: a coin toss or a quantum event (Widerker 1987). There is nothing that Peter is able to do to ensure that a coin lands heads, but Peter is able to do something such that a coin might land heads, e.g., he can flip the coin. The inevitability of all propositions – our inability to (strongly) falsify them – does not follow from the premises and assumptions of CA-3 because principle (β) is invalid given the strong sense of “render.”

There are two ways to fix CA-3 given the counterexample. The first is to switch to an *entailment version* of CA, which alters the transfer principle but keeps the operator the same, revising (β) to (β'), where “ $\Rightarrow$ ” stands for the entailment relation:

(β') From $Np, p \Rightarrow q$ deduce Nq .

Think of this as *strengthening the transfer principle* and it gives us the following argument (Widerker 1987; Finch and Warfield 1998):

(1') $(P_o \& L) \Rightarrow P$	assumption (df. ‘determinism’)
(3) $N(P_o \& L)$	assumption (necessity of the past & the laws)
(4) $\therefore N(P)$	from (1'), (3) by (β')

Premise (1') is intended to be equivalent to the former premise (1). The argument avoids the counterexample and as a bonus the corresponding version of MA presupposing indeterminism appears to be less plausible or at any rate very different.

Instead of strengthening the transfer principle, we could also weaken the operator. Though he began with the inevitability operator, van Inwagen (2000, 2015) eventually switches to an *unalterability* operator in an effort to respond to McKay and Johnson. Here is van Inwagen’s new operator:

‘ N^*p ’ stands for “ p and nothing that any human being is or ever has been able to do is such that if someone were to do it, that person’s action might result (could possibly result) in its not being the case that p ” (2015, p. 19; cf. Finch 2013, pp. 486–487).

Informally, “ N^*p ” means “it is a humanly unalterable truth that p ” (19). Here is a version of the *unalterability version* of CA-3.

(1) $\Box((P_o \& L) \supset F)$	assumption (df. ‘determinism’)
(2*) $N^*((P_o \& L) \supset F)$	from (1) by (α^*)
(3*) $N^*(P_o \& L)$	assumption (necessity of the past & the laws)
(4*) $N^*(F)$	from (2*), (3*) by (β^*)

The argument uses two inference rules:

- (α^*) From $\Box p$ deduce N^*p ;
 (β^*) From $N^*p, N^*(p \supset q)$ deduce N^*q .

Given (4*), all truths are humanly unalterable. Unalterability follows from determinism, and (leeway) free will skepticism follows from unalterability.

The worry about determinism was originally expressed in terms of inevitability but that approach ran afoul. Van Inwagen explains the transition to unalterability:

I mistakenly supposed that the only way in which it could be that one had no choice about the truth-value of a proposition would be for the truth-value of that proposition to be in some way so firmly “fixed” that one was unable to change it. I did not see that that there is another way for one to have no choice about the truth-value of a proposition: for that truth-value to be a mere matter of chance. (2015, p. 19)

This new worry – that determinism leads to the unalterability of all propositions – suggests an easier way to fix CA, one that is more in keeping with the results above. We merely replace the old operator with a new one.

There are two new versions of CA-3: the entailment version, which retains the inevitability operator, and the unalterability version. If we want to argue that CA-3 is equable to a version of MA, the entailment version is problematic, though Nelkin (2001) and Jenann Ismael (2016: Ch. 4) make strong cases.⁸ Not only does van Inwagen advocate the unalterability version, but there is also a more substantive reason for endorsing it: premise (1') is a red herring. The entailment in premise (1') is merely one way of conveying the tight connection between the past and the future given the laws of nature. Nomological relations remain, and are equally beyond the control of persons, given indeterminism. Showcasing this nomological relation as a relation of entailment misrepresents the metaphysical worry. What matters for the success of CA-3 is the supposition of nomological naturalism. This move is already needed to support premise (3*) but it can also be used to support proposition (2*), rendering premises like (1') superfluous:

(2*) $N^*((P_o \& L) \supset F)$	assumption (necessity of nomological relations)
(3*) $N^*(P_o \& L)$	assumption (necessity of the past & the laws)
(4*) $N^*(F)$	from (2*), (3*) by (β^*)

Several questions arise.

Mostly significantly, new support for proposition (2*) was motivated using the strong sense of “render,” but we have switched to the weak sense in the unalterability version of CA-3. Does the switch matter? The answer is tricky. We contend that premise (1') is gratuitous and misleading. There is an independent reason for endorsing proposition (2*): the unalterability of the nomological relations that underlie the laws of nature, whether they are deterministic or probabilistic. The reason why we cannot (strongly) falsify certain material conditionals is not that they are grounded by entailment relations between propositions but that they are grounded by nomological relations between events and by the applicable laws of nature.

On the other hand, our argument appeals to a closure principle like principle (β'). Suppose it is a law of nature that all A's are B's. Consider the material conditional:

- $a \text{ is } A \supset a \text{ is } B$

where a is an object that instantiates the law. If one is able to falsify the material conditional, one is able to falsify the law. Supposing laws of nature are propositions, the relevant relation between laws and material conditionals is entailment. In that case, our argument hinges on

the closure of unalterability and a principle like principle (β'), which expresses the closure of inevitability.

Transfer principles like (β') play a role in our reasoning about of CA-3 but not in a way that mirrors the reasoning of van Inwagen's original version of CA-3. The relevant entailment is not between propositions about the past and propositions about the future, but between the laws of nature and true material conditionals. The formulation of determinism as an entailment relation is an unnecessary step in the argument, so we prefer the unalterability version of CA-3, which retains the hope of finding an equitable version of MA.

4 Support of Premise (II): MA

Van Inwagen could have renamed Ch. III "One Argument for Incompatibilism Done Three Ways" (1983, p. 56) for he considers the versions of CA to be equable. "These three arguments, or versions of one argument," he continues, "are intended to support one another" (56). Things are much different for MA in the subsequent chapter.

Ch. IV is entitled "Three Arguments for Compatibilism" (1983, p. 106ff.) but there are only two arguments for that conclusion: the paradigm case argument (1983: §4.2) and the conditional analysis argument (§4.3). MA is the third argument considered (§4.4). It is called the "*Mind* argument" because influential versions of it were published in the journal of that name. A memorable example is "Free Will as Involving Determination, and Inconceivable Without It" (Hobart 1934) written by Dickinson S. Miller under the pseudonym R.E. Hobart. The title of the article conveys the radical spirit of this family of arguments.

MA "occurs in three forms" or "three closely related strands of argument that are often twisted together" (1983, p. 126). What these strands have in common is that they begin with "a certain set of reflections on what the nature of free action must be if the incompatibilist is right" (1983, p. 126), that is, supposing the world is similar to what incompatibilists say it is like. All the strands are critical of libertarianism, so they are in that sense supportive of compatibilism.

Following van Inwagen, we regard the three strands of MA as related but distinct arguments. Most importantly, their conclusions differ. The first strand of MA (MA-1; 1983, pp. 128–129) is sometimes called the *luck argument*, but that is a bad name given there are different kinds of luck. Indeed, the problem of free will and determinism is a version of the problem of *causal luck*, e.g., the worry that our actions might be causally determined by factors outside our control (Hartman Ch. 23). What MA-1 adds is that indeterminism also introduces luck, for then our actions are the result of random forces over which we have no control.

Elsewhere, van Inwagen (2000) gives the *rollback argument*, which highlights indeterminism, but randomness is not explicit. Imagine a situation in which a choice is made, then rollback time to some appropriate moment prior to the agent's decision. Keeping the past and the laws stable, how does the agent's action play out in subsequent recurrences? If we were to rollback time in a determined world, holding the past and the laws fixed, the same result would always follow. If determinism were false, different results might occur from the same same causes, a similar combination of reason and situation. Thus, reason does not appear to be the determining factor of one's actions in the case of indeterminism, and there is a worry whether reason is a factor at all. Such arguments have their detractors (Kane 1996).

The second strand (MA-2; 1983, pp. 129–142) draws a related but much stronger conclusion, getting its inspiration from the work of David Hume (1975, 1978). Here is a helpful passage from Miller:

I am not maintaining that determinism is true; only that it is true in so far as we have free will. ... But it is not here affirmed that there are no small exceptions, no slight undetermined swervings, no ingredient of absolute chance. All that is here said is that such absence of determination, if and so far as it exists, is no gain to freedom, but sheer loss of it; no advantage to the moral life, but blank subtraction from it. (Hobart 1934, p. 2)

As van Inwagen envisages it, MA-2 holds that “an undetermined act is a contradiction in terms” (1983, p. 134), and concludes that free willism “*entails* determinism” (1985, p. 40) for it requires a deterministic link between the agent – or features of the agent – and the action. Since Elizabeth Anscombe (1971) noted that indeterministic causes are possible, attitudes about causation being essentially deterministic have lost favor. Also, the conclusion is too strong for our needs, though we have more to say about Miller’s argument below.

Van Inwagen holds that “other statements of the [*Mind*] argument are available, including some that do not appeal to the concept of chance” (2000, p. 10). This is true of MA-3 (1983, pp. 142–152, 1985, 1989, 1994). He adds: “Though I’ve been treating this argument [MA-3] as an argument for the compatibility of free will and determinism, it is, strictly speaking, an argument for the *incompatibility* of free will and *indeterminism*” (1983, p. 148). He admits that the argument is problematic because “if it were sound, then either free will would be compatible with determinism or else free will would be incompatible with both determinism and indeterminism and hence impossible” (1983, p. 148). The argument serves our interests.

Part of the difficulty in understanding the similarities between CA-3 and MA-3 is that van Inwagen tells the motivating stories in different temporal directions. CA-3 is modeled on the informal version: The past and the laws are unalterable. It is unalterable that given the past and the laws the future follows. Ergo, nothing is alterable. Supposing indeterminism instead of determinism, telling the story from past to future doesn’t seem to work. It sounds contrived to suggest that indeterminism renders the future unalterable since indeterminism allows for more than one extension of the past and the laws. We forget that if indeterminism is true as we understand it – indeed for any indeterminism that has a chance of accommodating free will – the past is still nomically connected to the future, otherwise there is chaos and a disconnect between agent and act. At that point, MA-1 – the problem of indeterministic luck – becomes salient.

For this reason, van Inwagen presents MA-3 differently than CA-3, starting with an action and proceeding backwards to its ultimate cause. Here is the set-up.

Let us consider the case of a hardened thief who, as our story begins, is in the act of lifting the lid of the poor-box in a little country church. He sneers and curses when he sees what a pathetically small sum it contains. Still, business is business: he reaches for the money. Suddenly there flashes before his mind’s eye a picture of the face of his dying mother and he remembers the promise he made to her by her deathbed always to be honest and upright. This is not the first occasion on which he has had such a vision while performing some mean act of theft, but he has always disregarded it. This time, however, he does not disregard it. Instead, he thinks the matter over carefully and decides not to take the money. Acting on this decision he leaves the church empty handed. (1983, pp. 127–128)

Building toward MA-3, van Inwagen continues.

So let us assume at the outset that the thief had no choice about whether DB [“the indicated desire and belief”] occurred, for we should sooner or later have to make an assumption that would have the same philosophical consequence. But if the thief had no choice about whether DB occurred, and had no choice about whether, if DB occurred, then R [“refraining from robbing the poor-box”] followed, in that case, surely, the thief had no choice about whether R occurred. (1983, p. 146)

This sub-argument explicitly depends upon a transfer principle like (β^*), leaving it open that MA-3 and CA-3 are equable.

To draw the connection clearer, consider a semi-formal version of a sub-argument like CA-3.

- DB occurred, and the thief had no choice about whether DB occurred (DB is unalterable).
- If DB occurred, then R occurred, and the thief had no choice about whether if DB occurred, then R occurred (DB $\supset$ R is unalterable).
- $\therefore$ R occurred, and the thief had no choice about whether R occurred (R is unalterable).

Of course, one might insist that the thief had a choice about DB, yet we can expand the above reasoning and press the argument further.

Speaking generally, we start with some event or set of events E_1 (e.g., a mixture of the thief’s desires and beliefs) that is the cause of a decision D (e.g., the decision to not rob the poor-box) that is the cause of an act R (e.g., refraining from taking the money). Given nomological naturalism, event E_1 also has a cause, E_2 , which has another cause, E_3 , and so on. The significance is that even if indeterminism is true, we can walk back the causal chain of any action to a time in the individual’s remote past, a time prior to his birth, a time when nothing was up to him.

Van Inwagen’s argument depends on an assumption of *nomological waves*: the instant before an event occurs, the die is cast by the previous state of the world. If there are nomological waves between events, our actions might ultimately be causally connected to events that have faded into the remote past, ones over which no agent has or had any control. This is true even if the laws of nature are probabilistic. If we reverse the temporal order – from the past to the future – we get something equivalent to the core of CA-3, roughly:

(2*) $N^*((P_o \& L) \supset F)$	assumption (necessity of nomological relations)
(3*) $N^*(P_o \& L)$	assumption (necessity of the past & the laws)
(4*) $\therefore N^*(F)$	from (2*), (3*) by (β^*)

The remarks above suggest a defense of proposition (2*). Premise (3*) deserves the same level of support in both CA-3 and MA-3, for our choice about the past and the laws is the same whether determinism is true or not. (2*) and (3*) both rest in part on nomological naturalism, the view that there are lawlike connections between events in the past and events in the future. The central idea behind events having nomological waves is that one moment before their occurrence nothing is left for an agent to contribute. This provides a way of linking propositions about states of the world at a time to actions without the aid of determinism.

Thus far, the core of CA/MA consists of 3 logical parts: the necessity of the past and the laws; the necessity of nomological relations; and a no-choice transfer principle. Careful reflection reveals a redundancy since the first two parts both talk about laws of nature. Condensing the principles reveals *the core of CA/MA*.

- The unalterability of nomological relations;
- The unalterability of the remote past;
- No-choice transfer principles.⁹

All three elements of the core are determinism-independent.

5 A Determinism-Independent Argument for Free Will Skepticism

There are reasons to question the soundness of MA-3. Libertarians who hold that there is indeterminacy the moment of decision will likely reject proposition (2*). Recall in counterexamples to (2*) the presumed agent need not be able to *ensure* that F is false given P_o and L . She need only do something that *might* result in its being the case that F is false. Suppose Aidan has a 50–50 chance of deciding to raise her right hand at time t and she raises it. If there is a 50–50 chance of deciding whether she raises her hand or not, then it appears that Aidan *could have done* something that might have made the proposition – *that Aidan raises her hand at t* – false. Aidan could have falsified the proposition had she decided not to raise her hand, and that event had a 50–50 chance of occurring (cf. Kane 1996, pp. 107–115; pp. 142–143).¹⁰

The reply plays out well in the individual case but starts to break down when we calculate these probabilities across spacetime. First, there is Derk Pereboom’s “problem of wild coincidences” (1994, 2001, 2014). Suppose that each of a set of human actions is governed by 50–50 probabilistic laws. The difficulty comes in envisaging how the free actions performed by agents play out over the range of possible events given the laws. Pereboom notes that it is “too bizarre a scenario to believe, if for any substantial span of human history frequencies of indeterministically free choices should happen to dovetail with determinate physical probabilities” (1994, p. 30). In other words, it would be a wild coincidence that the actions decided by free, rational choices should all happen to line up exactly with worldly events dictated by general laws even if the laws are probabilistic.

It isn’t merely a matter of aligning the probabilities that govern types of free decisions with those that govern other types of worldly events. That is difficult enough but there are a multitude of interconnected probabilities at play. Say that a *time-slice* at t is a slice of four-dimensional spacetime containing only those events that comprise the state of the world at t . Hence, P_o is a proposition about a time-slice in the remote past. Besides the interconnections between worldly events that are orthogonal to given time-slices, noted by Pereboom, there are also nomological waves connecting the events of distinct time-slices. Van Inwagen makes a compelling case that there are nomological waves with regard to most if not all human decisions, which is why we have “precious little free will” (1989, p. 405) even if indeterminism is true. Pereboom extends the connections between events further by noting that our free actions must coordinate with other events that are subject to the same set of probabilistic laws, and we now extend them further still by exposing even more nomological waves involving events that are not actions but are merely parts of processes stretched over time. Events within and between different time-slices are wedded together by unavoidable nomological relations in a multitude of different ways.

The preceding passages suggest there is a very tight nomological connection between worldly time-slices. Based on this we endorse:

The Unalterability of Nomological Waves (UNW): If P_i and P_j are true propositions about time-slices of the world at instants i and j , respectively, and i occurs one moment prior to j , then $N(P_i \supset P_j)$.

UNW is the final piece needed for the determinism-independent argument for free will skepticism.

We prove no true propositions are alterable by mathematical induction. Using the core of CA/MA, we show that no true proposition about any time-slice is alterable, which entails the stronger conclusion. The *base case* concerns true propositions about the remote past. All true propositions about the remote past are unalterable since there were no humans around to alter them. If P_0 is a proposition about a time-slice in the remote past, P_0 is unalterable.

- N^*P_0 Necessity of the remote past

Next we need to establish the *induction step*: if a given true proposition about a time-slice is unalterable, then the true proposition about a time-slice at the next moment is unalterable. Suppose i and j are instants of time such that i occurs one moment prior to j . By hypothesis P_i is unalterable, but given UNW, P_j is also unalterable.

- N^*P_i Hypothesis for induction step
- $N^*(P_i \supset P_j)$ UNW
- $\therefore N^*P_j$ Principle (β^*)

This proves the induction step. Since the base case and induction step have been established, this completes the proof. True propositions about time-slices in the remote past are unalterable. Given that a true proposition about a time-slice is unalterable, the true proposition about the time-slice of the next moment is also unalterable. For any moment, no true proposition about a time-slice of the world at that moment is alterable, and (leeway) free will skepticism follows.

6 Concluding Remarks

The core of CA/MA is comprised of three determinism-independent parts:

- The unalterability of nomological relations;
- The unalterability of the remote past;
- No-choice transfer principles.

CA-3 and MA-3 are equable because there are versions of each argument with this common core.

Exactly what to conclude from the equability of CA-3 and MA-3 is unclear. Impossibilism seems to follow given our results, which is stronger than mere free will skepticism. Our lack of free will is independent of the specific nomological relations between our acts and earlier events. All that matters is that there are nomological relations. CA-3 does a good job of conveying the intuition that determinism compromises (leeway) free will, but precisely how this

implicates indeterminism needs further explanation. Similarly, the equability of CA-3 and MA-3 suggests that leeway free will is incoherent, a victory for source theorists. These are worth exploring, but each of these interpretations rests on the soundness of CA-3, and there is reason to question its soundness.

Our result places a burden on libertarian positions including van Inwagen's own brand of mysterianism, which rests on the inequity of CA-3 and MA-3. Perhaps one cannot detect a difference between them – after all, the mystery remains – but there is a difference, according to van Inwagen. On the other hand, if the arguments have the same core, and we are motivated to retain free willism despite them, we should endorse *compatibilist mysterianism*. If either argument is unsound, CA-3 is unsound, and if CA-3 is unsound, (leeway) incompatibilism loses its motivation.

Our thesis is *if CA-3 is sound, then so is MA-3*. This is a relative claim about the two arguments, and consistent with the failure of both. CA-3 makes assumptions that are too broad for its narrow target: leeway compatibilism. To the extent that it works, it takes down a wider range of theories than intended, e.g. all theories of leeway free will. If the core of CA-3 is too strong, establishing free will skepticism independently of the truth or falsity of determinism, this suggests a problem with the core. Being convinced of incompatibilism because of CA-3, you should also be convinced of free will skepticism. Alternatively, if you reject (leeway) free will skepticism, that provides a further reason to reject CA-3 as sound.¹¹

Notes

- 1 Arguments CA and MA support premises (I) and (II), respectively. Kristin Mickelson (Ch. 4) raises the worry that conditional claims of this kind are ambiguous (in the same way that many popular definitions of “incompatibilism” are ambiguous). On one interpretation, (I) asserts that there is a “perhaps spurious correlation relationship” between free willism and determinism, i.e., it just so happens the former is false whenever the latter is true. According to a second interpretation, (I) asserts that there is a “non-spurious relevance relation” between free willism and determinism, i.e. where free willism is false when determinism is true at least partly because determinism is true. As Mickelson explains, van Inwagen accepted background assumptions (e.g. possibilism and leeway free will) which ensured that any sound argument he gave for the former claim would, by what she calls “the classical bridge inference,” guarantee the truth of the latter. For this reason, we assume that CA, as van Inwagen initially presented it, aimed to establish both claims.
- 2 Thanks to Alan White for this point.
- 3 See White (Ch. 3) for support of local determinism based on his system-of-events local marker.
- 4 *Equability* is a relation between arguments similar to *incompatibility*, where two arguments are *incompatible* provided that “If either argument is sound (if its conclusion follows from its premises and if those premises are true), then the other is unsound” (van Inwagen 1985, p. 41). In 1985, van Inwagen argues that a version of MA is incompatible with the *Ethics* argument, aka the conditional analysis argument (1983: §4.3). See White (1990) for a reply. That CA-2 and CA-3 are equable is a controversial claim, as is the related contention that all arguments for incompatibilism are equable to transfer versions of CA. See van Inwagen (1994) and Fischer and Ravizza (1996). We maintain the equability of the formal versions of CA (Campbell 2017, pp. 158–161), which is why we use the expression *core of CA*.
- 5 One may ask how MA might compare to other arguments for incompatibilism, such as the manipulation argument (Capes Ch. 9). We think the comparison would produce similar results, but that is beyond the scope of this chapter.
- 6 Thanks to Fabio Lampert for pressing us about Warfield (2000). See Lampert (2022) for additional criticisms of the necessity of the past.

- 7 For an interesting discussion of Lehrer (1980) and Lewis (1981), see McKenna (2020).
 8 Here is a condensed version of Ismael's argument (2016, p. 98).
- Our best current physics entails that the facts of the past, in conjunction with the laws of nature and facts about the outcomes of chancy quantum processes, entail every truth about the future, including what we will do.
 - Neither the past, the laws of nature, nor the outcomes of chancy quantum processes are under our control (that is, they are unalterable).
 - Ergo our actions are not under our control (that is, they are unalterable).

The argument allows one to keep the entailment relation in spite of indeterminism by including "chancy quantum processes" among the items over which we lack control. One might contend that Ismael's argument runs together disparate kinds of luck, but it is worthy of further consideration.

- 9 Putting everything together suggests versions of CA-3 and MA-3 have this approximate structure:

$(2^{**}) N^*(P_o \supset F_n)$	assumption (unalterability of nomological relations)
$(3^{**}) N^*(P_o)$	assumption (necessity of the past)
$(4^{**}) \therefore N^*(F_n)$	from (2^*) , (3^*) by (β^*)

P_o and F_n are propositions about time-slices at times in the remote past and the future, respectively.

- 10 Thanks to Ish Haji for pressing us on this issue.
 11 Predecessors of this paper were delivered at the 2017 Illinois Philosophy Association Conference, the 2018 Northwest Philosophy Conference, Boise State University, the 2019 Rocky Mountain Ethics Conference, and the 2020 Science of Consciousness Conference. We thank participants in those audiences. For significant contributions, we thank Andrew Cortens, Zach Doering, Mylan Engle, John Martin Fischer, Alicia Finch, Bryan Haflich, Ish Haji, Ryan Hubert, Erik Inglis, Brian Kierland, Keith Lehrer, Dana Nelkin, Jacob Park, Jim Stone, Jake Wojtowicz, and especially Fabio Lampert, Kristin Mickelson, and Alan White.

Bibliography

- Anscombe, G.E.M. (1971). *Causality and Determination: An Inaugural Lecture*. London: Cambridge University Press.
- Bailey, A. (2012). Incompatibilism and the past. *Philosophy and Phenomenological Research* 85: 351–376.
- Beebe, H. and Mele, A. (2002). Humean compatibilism. *Mind* 111: 201–423.
- Campbell, J. (2007). Free will and the necessity of the past. *Analysis* 67: 105–111.
- Campbell, J. (2008). Reply to brueckner. *Analysis* 68: 264–269.
- Campbell, J. (2010). Incompatibilism and fatalism: reply to loss. *Analysis* 70: 71–76.
- Campbell, J. (2011). *Free Will*. Cambridge, UK: Polity Press.
- Campbell, J. (2013). Two problems for classical incompatibilism. In: *Free Will and Moral Responsibility* (ed. I. Haji and J. Couette). Newcastle upon Tyne: Cambridge Scholars Publishing.
- Campbell, J. (2017). The consequence argument. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy). New York, NY: Routledge.
- Finch, A. (2013). Against libertarianism. *Philosophical Studies* 166: 475–493.
- Finch, A. and Warfield, T. (1998). The mind argument and libertarianism. *Mind* 107: 515–528.
- Fischer, J.M. (1994). *The Metaphysics of Free Will: An Essay on Control*. Cambridge, MA: Blackwell Publishers.
- Fischer, J.M. and Ravizza, M. (1996). Free will and the modal principle. *Philosophical Studies* 83: 213–230.
- Ginet, C. (1966). Might we have no choice? In: Lehrer 1966.
- Ginet, C. (1990). *On Action*. Cambridge, UK: Cambridge University Press.

- Haji, I. (2009). *Incompatibilism's Allure: Principle Arguments for Incompatibilism*. Peterborough, Ont.: Broadview Press.
- Hobart, R.E. (Dickinson S. Miller) (1934). Free will as involving determination and inconceivable without it. *Mind* 43: 1–27.
- Hume, D. (1947). *Dialogues Concerning Natural Religion*, 2e (ed. N.K. Smith). Indianapolis: Bobbs-Merrill.
- Hume, D. (1975). *An Enquiry Concerning Human Understanding*. Oxford: Oxford University Press.
- Hume, D. (1978). *A Treatise of Human Nature*, 2e (ed. L.A. Selby-Bigge and P.H. Nidditch). Oxford: Oxford University Press.
- Ismael, J.T. (2016). *How Physics Makes Us Free*. Oxford: Oxford University Press.
- Kane, R. (1996). *The Significance of Free Will*. Oxford: Oxford University Press.
- Kneale, W. and Kneale, M. (1962). *The Development of Logic*. Oxford: Clarendon Press.
- Lampert, F. (2022). A puzzle about the fixity of the past. *Analysis* 82: 426–434.
- Lehrer, K. ed. (1966). *Freedom and Determinism*. New York: Random House.
- Lehrer, K. ed (1980). Preferences, conditionals and freedom. In: *Time and Cause: Essays Presented to Richard Taylor* (ed. V.I. Peter), 187–201. Dordrecht: D. Reidel.
- Lewis, D. (1981). Are we free to break the laws? *Theoria* 3: 113–121.
- Lewis, D. (1986). Introduction. *Philosophical Papers, II*. Oxford: Oxford University Press.
- Long, A.A. and Sedley, D.N. eds. (1987). *The Hellenistic Philosophers*, 1. Cambridge: Cambridge University Press.
- McKay, T. and Johnson, D. (1996). A reconsideration of an argument against compatibilism. *Philosophical Topics* 24: 113–122.
- McKenna, M. (2020). A lost lesson in Keith Lehrer's reply to the consequence argument. *Grazer Philosophische Studien* 97: 545–558.
- Nelkin, D. (2001). The consequence argument and the *mind* argument. *Analysis* 61: 107–115.
- Pereboom, D. (1994). Determinism *al Dente*. *Nous* 29: 21–45.
- Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
- Perry, J. (2004). Compatibilist options. In: *Freedom and Determinism* (ed. J.K. Campbell, M. O'Rourke and D. Shier). MIT Press.
- Sartorio, C. (2016). *Causation and Free Will*. Oxford: Oxford University Press.
- Sellars, W. (1966). Fatalism and freedom. In: Lehrer 1966.
- van Inwagen, P. (1974). A formal approach to the problem of free will and determinism. *Theoria: A Swedish Journal of Philosophy* 40: 9–22.
- van Inwagen, P. (1975). The incompatibility of free will and determinism. *Philosophical Studies* 27: 185–199.
- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.
- van Inwagen, P. (1985). On two arguments for compatibilism. *Analysis* 45: 161–163. Reprinted in van Inwagen 2017. Page numbers are to the 2017 version.
- van Inwagen, P. (1989). When is the will free? *Philosophical Perspectives* 3: 399–422. Reprinted in van Inwagen 2017.
- van Inwagen, P. (1994). When the will is not free. *Philosophical Studies* 75: 95–113. Reprinted in van Inwagen 2017.
- van Inwagen, P. (2000). Free will remains a mystery. *Philosophical Perspectives* 14: 1–19. Reprinted in Kane 2002. Reprinted in van Inwagen 2017.
- van Inwagen, P. (2004). Van Inwagen on free will. In: *Freedom and Determinism* (ed. J.K. Campbell, M. O'Rourke and D. Shier). Cambridge, MA: MIT Press.
- van Inwagen, P. (2015). Some thoughts on an essay on free will. *Harvard Review of Philosophy* 22: 16–30.
- van Inwagen, P. (2017). *Thinking about Free Will*. Cambridge: Cambridge University Press.
- Vihvelin, K. (2013). *Causes, Laws, & Free Will*. Oxford: Oxford University Press.

- Warfield, T.A. (2000). Causal determinism and human freedom are incompatible: a new argument for incompatibilism. *Philosophical Perspectives* 14: 167–180.
- White, V.A. (1990). How to *mind* one's *ethics*: a reply to van Inwagen. *Analysis* 50: 33–35.
- Widerker, D. (1987). On an argument for incompatibilism. *Analysis* 47: 37–41.
- Zhang, J. (2013). Can the incompatibilist get past the no past objection? *Dialectica* 67: 345–352.

Manipulation and Direct Arguments

JUSTIN A. CAPES

Broad incompatibilism is the thesis that both free will and moral responsibility are incompatible with determinism.¹ Arguments for this thesis commonly rely on two main ideas. The first is that free will and moral responsibility require avoidability – to act freely and be morally responsible for your behavior and its consequences, you must sometimes have it within your power to avoid these things, to prevent them from occurring. The second main idea is that the requisite power of avoidability is incompatible with determinism. Neither of these ideas is uncontroversial, though. Indeed, the debates surrounding them are among the most intricate and seemingly intractable in the literature on free will and moral responsibility. This fact partly explains why, in recent years, philosophers have become increasingly interested in arguments for incompatibilism that don't rely on those two ideas. The arguments I have in mind are also interesting in their own right. They tend to fall into two main categories: manipulation arguments and direct arguments.²

1 Manipulation Arguments

Consider manipulation arguments first. These arguments take different forms, but the main thing they have in common, the thing that makes them “manipulation” arguments, is that they are all based around science fiction vignettes in which a controller – a meddling deity, for example, or a sophisticated neuroscientist or behavioral engineer – tinkers with an agent in a way that ensures that the agent does what the controller wants him to do. Different manipulation arguments appeal to different vignettes of this sort and to different premises in arguing from judgments about the vignettes to the conclusion that determinism precludes free will and moral responsibility.³

A basic manipulation argument begins with a relatively straightforward case of manipulation in which a powerful controller alters an agent's mental states more or less directly (e.g., by using a mind-control device that has been covertly installed in the agent's brain), thereby ensuring that the agent performs an action *A* that the agent almost certainly wouldn't have performed otherwise. The basic manipulation argument then proceeds as follows:

A Companion to Free Will, First Edition. Edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White.
© 2023 John Wiley & Sons, Inc. Published 2023 by John Wiley & Sons, Inc.

1. *No-responsibility-premise*: the manipulated agent in this case didn't *A* of his own free will and so isn't morally responsible for *A*-ing.
2. *No-difference-premise*: when it comes to free will and moral responsibility, there is no relevant difference between the manipulated agent in this case and ordinary, non-manipulated agents in fully deterministic settings.
3. Hence, agents in fully deterministic settings don't act freely and so aren't morally responsible for their behavior.

The argument is valid, but how plausible are its premises?⁴ This will depend in large part on the details of the vignette. Some forms of manipulation are clearly inconsistent with the satisfaction of uncontroversial requirements for free will and moral responsibility. If the agent was manipulated in one of those ways, then the first premise will be true, but the second premise will be false, since ordinary agents in deterministic settings can satisfy the requirements in question.

To take a simple illustration, suppose the manipulation resulted in the agent becoming completely insane, which in turn ensured that the agent would *A*. In that case, while the no-responsibility-premise of the basic manipulation argument would clearly be true, the no-difference-premise would just as clearly be false. Because some level of sanity is plausibly required for free will and moral responsibility, and because determinism is consistent with the requisite level of sanity, there is a relevant difference between the manipulated agent in this case and ordinary agents in deterministic settings. Whereas the manipulated agent doesn't satisfy the sanity requirement, ordinary agents in deterministic settings definitely would satisfy it.

What this simple case illustrates is that to make a manipulation argument at all plausible the scenario on which it's based can't involve just any form of manipulation. The manipulation has to be such that the agent still satisfies mundane requirements for free will and moral responsibility of the sort that can easily be satisfied if determinism is true. The manipulation must leave the agent sufficiently sane, rational, responsive to reasons, etc. However, the case must also be one in which the agent arguably doesn't act freely and isn't morally responsible for his behavior. The trick, then, is to find a manipulation scenario in which the agent satisfies standard compatibilist conditions for free will and moral responsibility while seemingly lacking the kind of control over his behavior required for him to be morally responsible for it.

Bearing this in mind, consider a manipulation story of Derk Pereboom's in which Professor Plum decides to murder Ms. White for selfish reasons. We are to imagine that "Plum is just like an ordinary human being, except that a team of neuroscientists programmed him at the beginning of his life so that" he often reasons in selfish ways. The programming has the intended result that in the present situation Plum "is causally determined to engage in the egoistic reasons-responsive process of deliberation and to have the set of ... desires that result in his decision to kill White." The manipulation, in other words, ensures that Plum will decide to kill White, but it does so without rendering Plum insane or otherwise irrational (2014, p. 77).⁵

Many will have the intuition that, in this case, Plum didn't kill Ms. White freely and isn't morally responsible for killing her. After all, he was programmed to do just that. He was, in effect, the puppet of the neuroscientist. But, unlike the simple manipulation case we considered above, the no-difference claim can't be so easily rejected this time, since Plum clearly satisfies a variety of uncontroversial requirements for free will and moral responsibility – he's sane, rational, etc. Indeed, as Pereboom constructs the story, the neuroscientists tinker with Plum in a way that ensures he will make the decision to kill Ms. White and that, in doing so,

he will satisfy sufficient conditions for free will and moral responsibility specified by leading compatibilist accounts of these things. Given that Plum satisfies those requirements, it isn't clear whether there is a relevant difference between him and an ordinary agent in a deterministic setting.⁶ If there isn't any such difference, and if Plum is indeed off the hook for deciding to kill Ms. White, then the leading compatibilists accounts of free will and moral responsibility are unsuccessful and a compelling manipulation argument for broad incompatibilism would seem to be in the offing.

Consider in this connection a story of Alfred Mele's. In it, a powerful goddess, Diana, creates a zygote in Mary, which she designs specifically so that it will develop into an otherwise ordinary human being, Ernie, who, thirty years later, will bring about an event that Diana very much wants to occur – the death of Ernie's aunt. Diana arranges the atoms of the zygote in such a way that, given the laws of nature and the state of the universe at the time, ensures that Ernie will develop into an otherwise normal agent who will bring about the death of his aunt at the appointed time by poisoning her. Diana's plan succeeds, as she knew it would. "Thirty years later," at the time of the murder, "Ernie is a mentally healthy, ideally self-controlled person who regularly exercise his powers of self-control and has no relevant compelled or coercively produced attitudes. Further, his beliefs are conducive to informed deliberation about all matters that concern him, and he is a reliable deliberator" (2006, p. 188). In short, at the time of the murder, Ernie satisfies a robust set of compatibilist conditions for free will and moral responsibility.

Mele notes that broad incompatibilists could use the story about Ernie as the basis for the following argument for their position. Mele calls it "the zygote argument."

1. Because of the way his zygote was produced in his deterministic universe, Ernie is not a free agent and is not morally responsible for anything.
2. Concerning free action and moral responsibility of the beings into whom the zygotes develop, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe.
3. So determinism precludes free action and moral responsibility (2006, p. 189).⁷

The zygote argument has a similar structure to the basic manipulation argument introduced earlier. Like that argument, it has a no-responsibility premise and a no-difference premise. The primary difference between the two arguments is the vignettes on which they are based. Rather than appealing to a case in which a controller manipulates an already existing agent, the zygote argument begins with a story in which the controller, Diana, creates an agent from scratch, building in from the very beginning the necessary physiological and psychological ingredients necessary to ensure that the agent will behave at some later time in the way the controller wants him to behave at that time. Given this difference, Mele remarks that the "story about Ernie is not a story about manipulation, and the associated original-design argument [i.e., the zygote argument]...is not a manipulation argument" (2008, p. 285). But because there seem to be more structural similarities than differences between basic manipulation arguments and original design arguments like this one, the latter are often numbered among the former.

Objections to manipulation (and original design) arguments come in two main varieties.⁸ Michael McKenna (2008a) refers to them as *hard-line* and *soft-line* replies, respectively. Hard-line replies target the no-responsibility-premise by arguing that, as long as the manipulation doesn't prevent the agent from satisfying certain crucial conditions for free will and moral responsibility, conditions the satisfaction of which are compatible with determinism, the

agent may well have acted freely and may be morally responsible for what he did, the fact that he was “set up” by the controller notwithstanding. Soft-line replies typically grant the no-responsibility premise and instead challenge the no-difference premises by trying to identify relevant differences between agents in the various manipulation scenarios and ordinary, non-manipulated agents in deterministic settings, differences that would explain why the manipulated agent doesn’t act freely and isn’t morally responsible, but which wouldn’t imply that ordinary agents in deterministic settings lack free will and moral responsibility.

The main worry people have with the hard-line approach is that it conflicts too deeply with our intuitions about the relevant class of manipulation (and original design) cases. Many people find it highly implausible that someone like Plum or Ernie could be genuinely responsible for his or her behavior given how the behavior came about. Proponents of the hard-line approach have suggested several strategies for addressing this worry, but their main strategy is to challenge the reliability of our intuitions about suitably constructed manipulation cases. John Fischer (2011) and McKenna (2008a, 2014), for example, agree that, when we begin with manipulation stories like the ones involving Plum or Ernie, it can seem that the agent in those cases isn’t responsible and doesn’t act freely. However, Fischer and McKenna go on to argue that, when we begin with ordinary cases of causal determinism, many people will have the opposite reaction, judging the agent to be responsible and to have acted freely. And, according to McKenna (2008a, p. 157, 2014, pp. 480–481), we should give precedence to our intuitions about ordinary cases. But now, if there really is no freedom or responsibility relevant difference between agents in an ordinary deterministic settings and agents in suitably developed manipulation (or original design) stories, then we should judge that agents in those stories act freely and are responsible for their behavior, the actions of the controller notwithstanding.

Carolina Sartorio (2016) has recently augmented the preceding strategy by suggesting an error theory of sorts designed to explain why we might initially have the feeling that manipulated agents like Plum and Ernie don’t act freely and aren’t morally responsible despite satisfying the relevant compatibilist conditions. Her suggestion is that, in considering manipulation stories like those proffered by Pereboom and Mele, there is a “dilution of control” effect wherein the control exercised by the manipulator is more noticeable than the control exercised by the manipulated (or designed) agent, leading us to make an erroneous judgment that the agent lacks sufficient control over his behavior to be responsible for it. “Perhaps,” Sartorio writes, “given the salience of Diana’s contribution [in Mele’s example], we tend to see Diana as more in control of Ernie’s acts than Ernie himself, when in fact this is a misguided perception” (169). If Sartorio is correct that this psychological effect is what accounts for the intuition that agents like Plum and Ernie don’t act freely and aren’t responsible for their behavior, it would cast doubt on the intuitive plausibility of the no-responsibility premises of leading manipulation arguments.

Hannah Tierney (2013, 2014) has developed a different variation of the hard-line reply that involves adopting a plausible scalar view of moral responsibility, according to which agents deserve more or less praise or blame depending on the circumstances. Though not myself a compatibilist, I have independently developed a similar position (Capes 2013). Both Tierney and I claim that, provided an agent like Plum or Ernie satisfies the relevant compatibilist conditions for free agency and moral responsibility, compatibilists should insist that the agent did act freely and that he bears at least some responsibility for his action, but that his responsibility is (or may be) diminished in that he deserves less praise or blame, less punishment or reward, for what he did than he would have if his action hadn’t been causally determined in the way it was.

This variation on the hard-line reply has several advantages. First, it fits nicely with the independently plausible view that responsibility comes in degrees, that agents can deserve more or less praise or blame depending on whether mitigating or aggravating circumstances obtain. Second, as Tierney observes, the reply also fits nicely with the fact that, in real life cases, considerations about the circumstances that led people to behave as they do, especially their formative circumstances, often lead to a reduction in the amount of blame or punishment we think is appropriate, without leading us to let the person off the hook completely.⁹ Third, and perhaps most importantly, this variation of the hard-line reply allows compatibilists to acknowledge, to some extent at least, the intuition that an agent's having been causally determined by factors beyond his control has an impact on his moral responsibility, but it allows them to do so in a way that doesn't force them to abandon the core compatibilist position.

Despite these advantages, this variation on the hard-line reply faces a challenge. Patrick Todd puts the challenge this way: "if the compatibilist admits that determinism itself is mitigating, a fair question is, in virtue of what?" (2011, p. 131). Tierney and I respond to Todd's challenge in different ways. Tierney suggests that the conditions for responsibility might come in degrees and that agents like Plum or Ernie might satisfy these requirements to a lesser degree owing to how their behavior was brought about. One might, for example, be more or less responsive to reasons, and the causal history of a person's action might result in the person being less responsive to reasons, and thus deserving of less praise or blame, than he would have been had his action been caused in a different way. I contend that determinism might preclude things like the power to do otherwise or being the ultimate source of one's actions that are required for full or maximal responsibility, but which, it could be argued, aren't required for a minimal level of responsibility, a level compatible with the truth of determinism. Whether either response successfully addresses the challenge Todd poses is an issue that merits further investigation.¹⁰

Let's turn now to soft-line replies to manipulation arguments, which, you'll recall, typically grant that manipulated agents like Plum and Ernie aren't responsible, while challenging the claim that there is no freedom or responsibility relevant difference between agents like that and ordinary, non-manipulated agents in deterministic settings. The main worry about the soft-line approach is that it will turn out to be merely a stopgap.¹¹ Even if proponents of the approach can point to relevant differences between the agent in the particular manipulation story under consideration and ordinary agents in deterministic settings, this might seem like little more than an invitation for proponents of the manipulation argument to revise the vignette on which their argument is based so that the agent now satisfies the relevant conditions on which the soft-liner insists. The challenge, then, for defenders of the soft-line approach is to show that there is a requirement for free action and moral responsibility that can be satisfied by agents in deterministic settings but which can't be satisfied by agents in manipulation (or original design) cases. Unless proponents of the soft-line approach can answer this challenge, their objections won't be sufficient to disarm the most carefully crafted manipulation arguments.¹²

There have been several attempts by proponents of the soft-line approach to address this challenge. Here I'll briefly mention one. In a recent article, Oisín Deery and Eddy Nahmias (2017), claim that, to act freely and be morally responsible, an agent must be, in some sense, the causal source of her behavior. In this they agree with many incompatibilists. However, they go on to argue that an agent can be the causal source of her action, in the relevant sense, even if determinism is true, since nothing beyond the agent's control "will have a stronger causal-explanatory relationship with her actions than relevant variables with her

control" (2). But in the case of manipulated agents like Plum and Ernie, Deery and Nahmias contend that things are importantly different, for in those cases, they argue, something beyond the agent's control has a stronger causal-explanatory relationship with the agent's actions, viz., the intentions and activity of the controller. If Deery and Nahmias are right about this, then there is a relevant difference between manipulated agents and ordinary agents in deterministic settings. Whether this or other attempts to defend the soft-line position survive critical scrutiny remains to be seen.

2 Direct Arguments

Let's turn now to the direct arguments, so named because they aim to establish the incompatibility of determinism and moral responsibility "directly," i.e., without first establishing the intermediary conclusion that determinism precludes the power to do otherwise. One thing to keep in mind about these arguments is that, strictly speaking, they only purport to establish incompatibilism about determinism and responsibility. But, as we'll see, if they succeed in establishing that conclusion, they can arguably be augmented to support incompatibilism about free will as well. For this reason, they count as arguments for broad incompatibilism.

Whereas the defining feature of manipulations arguments is a vignette involving behavioral controllers, the defining feature of direct arguments is some sort of transfer of non-responsibility principle. Principles of this sort purport to identify conditions in which lack of responsibility for one set of circumstances is transmitted to its consequences. The basic idea behind direct arguments is that our lack of responsibility for events in the distant past transfers to the inevitable effects of those events, which, if determinism is true, includes all of our behavior.

The original direct argument is due to Peter van Inwagen (1980, 1983, pp. 182–188). Here's a slightly simplified version of the argument, where L stands for the laws of nature, H for a complete description of the state of the universe a billion years ago, and P is any fact:

1. Determinism is true [assumption for conditional proof].
2. Necessarily, if (H & L), then P [consequence of 1].
3. Necessarily, no one is even partly responsible for necessary truths [premise].
4. So, necessarily, no one is even partly responsible for the fact that if (H & L), then P [from 2 and 3]
5. No one is even partly responsible for the fact that (H & L) [premise].¹³
6. Necessarily, if no one is even partly responsible for the fact that *if p then q*, and if no one is even partly responsible for the fact that *p*, then no one is even partly responsible for the fact that *q* [transfer of non-responsibility premise].
7. So, no one is even partly responsible for the fact that P [from 4, 5, and 6].

If sound, this argument shows that no one is even partly responsible for any truth, if determinism is true. But is it sound? As the argument is reconstructed here, it's clearly valid. The question, then, is whether its premises (steps 3, 5, and 6) are true. Let's consider each premise in turn.

The argument's first premise is step 3, the claim that, necessarily, no one is even partly responsible for necessary truths. This premise went unchallenged for years, and understandably so. Initially, it's extremely difficult to see how contingent agents like you and I could

be responsible for metaphysically necessary truths. Such truths are, after all, true independently of us – or at any rate, so it would seem. These truths were true long before any of us was born, will be true no matter what any of us do or fail to do, and, indeed, would be true even if none of us had ever existed. How, then, could anyone be even partly responsible for them, given that their truth seems so utterly independent of us and anything we have done or could ever do?

In recent years, though, 3 has come under increased scrutiny. Both Stephen Kearns (2011) and Charles Hermes (2014) object to it by challenging, albeit in different ways, the claim that all necessary truths are completely independent of human behavior. Both Kearns and Hermes try to show that there are necessary truths the truth of which is due, in some sense, to the behavior of a contingent agent and so is something for which that agent might be responsible. Here's an example of Kearns's: "It is necessarily true that [the song] Hey Jude starts in the distinctive way it does." But, according to Kearns, this fact (the fact that Hey Jude starts the way it does) is something for which Paul McCartney is responsible, since it was McCartney who composed the song and who made it the case that the song begins in that way. Kearns concludes that McCartney is responsible for a necessary truth and thus that 3 is false (Kearns 2011, p. 312).

Three responses to this sort of objection are worth mentioning, though space limitations prohibit a detailed assessment of them. The first is to argue that the necessary truths to which Kearns and Hermes appeal are not in fact true because of anything contingent agents have done or failed to do, at least not in any sense of "because" that renders these agents even partly responsible for those truths.¹⁴ The second is to concede that 3 is false, but to develop a direct argument that doesn't rely on the principle. If this strategy could be successfully executed, then while the objections of Kearns and Hermes might undermine some direct arguments, they wouldn't threaten all of them.¹⁵ The third response also concedes that 3, as currently stated, is false, but contends that it simply needs to be restricted in the following way: necessarily, no one is even partly responsible for necessary truths the truth of which isn't due, in the relevant sense, to the activity of human agents. Proponents of this third response could then argue that the necessary truths at issue in direct arguments aren't grounded in the activity of human agents, in which case the revised principle can still be used as part of an argument of that type. Revising the principle in this way might initially seem ad hoc, but it does have the advantage of explaining why some necessary truths – that $49 \times 18 = 882$, for example – seem to be beyond the realm of human responsibility, whereas others, like those cited by Kearns and Hermes, may not be. Whether any of these three replies can be developed successfully is a matter for further research.

The argument's next premise is step 5, the claim that no one is even partly responsible for the state of the universe a billion years ago and the laws of nature. No one disputes this premise.¹⁶ However, David Widerker (2002) suggests that while the premise is no doubt true, it nevertheless proves to be problematic for proponents of the direct argument. How so? Because the only motivation for it, he claims, appeals to the idea that we don't have the power to prevent the past or the laws of nature and so can't be responsible for them. But recall that direct arguments were supposed to establish the incompatibility of responsibility and determinism without appealing to ideas like that. Hence, Widerker concludes that direct arguments don't have the dialectical advantage they initially seemed to have, since they do ultimately rely on claims about avoidability and the power to do otherwise.

Is Widerker right about this? I'm not convinced that he is. Consider the following defense of 5:

- 5a. If x isn't an action, omission, or a result of someone's action or omission, then no one is even partly morally responsible for x .
 5b. (H & L) isn't an action, omission, or a result of someone's action or omission.
 5. Hence, no one is even partly responsible for the fact that (H & L).

This defense of 5 is plausible. Moreover, it appears not to rely on claims about avoidability, and although I can't pursue the issue here, I'm fairly confident that this appearance isn't illusory.

The direct argument's third and final premise is step 6, which, you'll recall, states that, necessarily, if no one is even partly responsible for the fact that *if p then q* , and if no one is even partly responsible for the fact that p , then no one is even partly responsible for the fact that q . This is the transfer of non-responsibility principle on which van Inwagen's direct argument turns, and it's the step of the argument that has received the lion's share of critical attention, most of which is devoted to identifying counterexamples to the principle. The examples to which critics of the principle have appealed vary in structure, but, in my estimation, the most worrisome cases for the principle are those involving some sort of overdetermination.¹⁷

Consider, for instance, the following scenario:

Kathy stands to gain a substantial amount of money if her husband, Earl, dies before noon. So, at 11:49, she injects Earl with a lethal poison for which there is no antidote with the intention of killing him before noon. The poison with which she injected Earl takes no longer than ten minutes to kill. So, by giving him the poison at 11:49, Kathy ensures that Earl meets his untimely end by 11:59, at the latest. Kathy does all this of her own free will, despite knowing that it is morally wrong. Five minutes later, at 11:54, Earl ingests an even deadlier substance, one that takes no more than five minutes to kill. His ingesting this second substance at 11:54 is sufficient, given the laws of nature and other background conditions, to bring about his death before noon. Both the poison given to Earl by Kathy and this second substance contribute to Earl's demise at 11:59. (Capes 2016, p. 1480)

This story arguably yields a counterexample to van Inwagen's transfer of non-responsibility principle.¹⁸ No one is even partly responsible for the fact that *if Earl ingests the second poison at 11:54, then Earl dies before noon*. Nor, we may suppose, is anyone even partly responsible for the fact that *Earl ingests the second poison at 11:54*. Contrary to what van Inwagen's principle implies, however, it doesn't follow that no one is even partly responsible for the fact that *Earl dies before noon*. Kathy is arguably at least partly to blame for that fact. It was, after all, her free action of poisoning him that first ensured that he would die before noon. Hence, it seems that van Inwagen's transfer principle is false, in which case his direct argument is unsound.

While many people agree with this assessment of the situation, there is widespread disagreement about how serious a problem it is for the general strategy underlying direct arguments. Some have suggested that adding a few comparatively minor qualifications to van Inwagen's transfer of non-responsibility principle will render it immune to counterexamples like the one just cited.¹⁹ Others have suggested that there are different transfer of non-responsibility principles in the neighborhood that avoid the difficulties facing van Inwagen's and which can be used to develop other direct arguments for the incompatibility of responsibility and determinism.

McKenna (2001, 2008b, p. 364), for instance, though not himself an incompatibilist, suggests that a transfer of non-responsibility principle restricted to cases in which there is only one causal sequence leading to the relevant outcome could be used to generate an initially compelling direct argument.²⁰ The principle McKenna suggests is rather complex, but the details needn't concern us. The important thing to note for present purposes is that a transfer principle restricted in this way isn't threatened by the story about Kathy and Earl,

since, in that story, there are two independent causal sequences (the one involving Kathy's murderous act and the one involving the second deadly substance) leading to the relevant outcome (Earl's death at 11:59).

In Capes (2016), I also suggest a transfer of non-responsibility principle that isn't threatened by overdetermination cases and which can be used to generate a direct argument for the incompatibility of responsibility and determinism. The principle says, roughly, that if a person isn't responsible for any of the circumstances that led to a particular outcome, and if that person isn't responsible for the fact that those circumstances led to that particular outcome, then the person isn't responsible for the outcome in question either. The story about Kathy and Earl doesn't pose a problem for this principle either. For the story to threaten the principle, it would have to be one in which Kathy isn't responsible for *any* of the circumstances that led to Earl's dying before noon. But she is responsible for *some* of those circumstances; she's responsible for poisoning Earl, an action that ensured that he would die prior to noon. Hence, overdetermination cases like this don't have the right structure to threaten this particular principle, nor, as far as I can tell, do any of the other alleged counterexamples to 6 that have appeared in the literature.

Even if proponents of direct arguments can identify a transfer of non-responsibility principle that is immune to obvious counterexamples, there is another hurdle they must overcome. Michael McKenna (2008b) suggests that, in order to show that the relevant transfer principle is true, proponents of the principle must show that it works in ordinary cases of human agency, cases in which the agent's behavior is caused in a normal way, since it's those cases (at least when they occur in fully deterministic settings) about which compatibilists and incompatibilists disagree. However, as McKenna goes on to point out, the sorts of examples typically used to motivate the various transfer of non-responsibility principles aren't of that sort. McKenna concludes that proponents of the various direct arguments have failed to motivate the key premise of those arguments.²¹ In my estimation, motivating the relevant transfer principle is the most pressing challenge facing proponents of the various direct arguments.

One final issue concerning direct arguments merits attention. As I mentioned earlier, these arguments only purport to establish the incompatibility of responsibility and determinism. For this reason, it might seem that they are irrelevant to the question of whether determinism is compatible with free will. But this seeming is illusory. If any direct argument succeeds in establishing the incompatibility of responsibility and determinism, it can arguably be extended to show that determinism precludes free will, too. To do so, one only needs to add the following premise: if free will is compatible with determinism, then so is responsibility. This premise, when conjoined with the conclusion of a direct argument that responsibility is *not* compatible with determinism, entails that free will isn't compatible with determinism either.

How plausible is the additional premise, though? Here's an argument for it.²² Suppose that free will is compatible with determinism. Now imagine an ordinary agent in a deterministic universe who freely performs some immoral action *A* and who satisfies other relatively mundane requirements for free will and moral responsibility (e.g., the agent is sane, responsive to reasons, he knew at the time that *A*-ing then was wrong, and so on). Given these assumptions, it's quite plausible that the agent is morally responsible (and blameworthy, in particular) for *A*-ing. After all, he *A*-ed of his own free will, knew perfectly well that *A*-ing was wrong, and satisfied all other standard conditions for moral responsibility. But if he is morally responsible for *A*-ing, the truth of determinism notwithstanding, then responsibility and determinism are compatible. So, if free will is compatible with determinism, it seems that moral responsibility must be as well.

With this additional premise in hand, the direct case for broad incompatibilism is complete. Here's a brief summary of it:

1. If free will is compatible with determinism, then so is responsibility (additional premise).
2. Responsibility isn't compatible with determinism (result of a direct argument).
3. So, free will isn't compatible with determinism either.

3 Conclusion

Manipulation and direct arguments are interesting for a number of reasons, not least of which is that they endeavor to establish broad incompatibilism without appealing to controversial claims about avoidability and its relationship to determinism, free will, and moral responsibility. Whether these arguments succeed in establishing the incompatibilist conclusion, and whether they do so without appealing to claims about avoidability, is, as we have seen, a matter of some dispute. But if current trends are any indication, such arguments will figure prominently in the debate between broad incompatibilists and their critics for the foreseeable future.²³

Notes

- 1 The term "broad incompatibilism" is due to Clarke (2003, p. 12). "Determinism" is, roughly, "The thesis that there is at any instant exactly one physically possible future" (van Inwagen 1983, p. 3). There is no agreed-upon definition of the term "free will," but the two most common are (1) the power or ability to voluntarily do otherwise (see, e.g., van Inwagen 1983, p. 8), and (2) the strongest sort of agential control required for moral responsibility (see, e.g., Mele 2006, p. 17). "Moral responsibility" refers, roughly, to deserving praise or blame.
- 2 Another kind of argument for broad incompatibilism that doesn't explicitly rely on premises about avoidability but which won't be discussed in any detail here are "ultimacy" arguments. At the heart of these arguments is the idea that to have free will and be morally responsible a person must be in some sense, some sense that's incompatible with determinism, the ultimate source or originator of his action. For a recent critical discussion of such arguments, including a list of references, see Sartorio (2016, pp. 148–153). I should note, however, that manipulation arguments might simply be one form of ultimacy argument. On one way of looking at manipulation arguments, they are designed to support the view that the sort of sourcehood or self-determination required for free and responsible agency is incompatible with determinism. For this view of the matter, see Tognazzini (2014).
- 3 See Todd (2017) for a discussion of the relationship between manipulation arguments and arguments for broad incompatibilism that rely on claims about avoidability.
- 4 Mickelson (2015) contends that arguments like this are invalid. But that's because she defines "incompatibilism" differently than I've defined it here. As I understand it, to say that two things are incompatible is to say that you can't have both simultaneously. So, to say that free will and moral responsibility are incompatible with determinism is simply to say that no one can have free will or be morally responsible if determinism is true. When "incompatible" is understood in this way, this basic manipulation argument is clearly valid. For further discussion of the issue Mickelson raises in her paper, see De Marco (2016).
- 5 An earlier version of this story appears in Pereboom (2001, pp. 113–114).
- 6 Pereboom often develops his manipulation argument by appealing to a best explanation premise rather than a no-difference premise. He claims, for example, that the best explanation of Plum's

- lack of moral responsibility in this case is that Plum was causally determined to behave as he did by factors beyond his control (see, e.g., Pereboom 2001, p. 114, 2014, p. 78). I omit here discussion of this claim. For a criticism of it, see Mele (2005, 2006, pp. 138–144, 2008, pp. 276–278). For a reply to Mele’s criticism, see Pereboom (2014, pp. 82–89).
- 7 Mele himself isn’t an incompatibilist and doesn’t endorse this argument, but he says that it has enough intuitive pull for him to prevent him “from flatly endorsing compatibilism” (2006, p. 193).
 - 8 For an objection to the zygote argument that doesn’t fall neatly into either the soft or hard line category, see Kearns (2012). For a reply to Kearns, see Todd (2013).
 - 9 She cites Clarence Darrow’s defense of Leopold and Loeb in this connection. One also thinks of Gary Watson’s (1987) influential discussion of the murder Robert Harris. When we first hear the details of Harris’s crime and his callous response to it, we are strongly inclined to hold him responsible for what he did and, in particular, to think that a pretty severe punishment is warranted. But when we subsequently learn about the unimaginable emotional and physical abuse Harris suffered as a child, our outrage subsides, at least to some extent. One possible interpretation of our reaction is that we recognize that Harris’s terrible upbringing somehow contributed to his equally terrible behavior in a way that diminishes his responsibility for his crime without rendering him totally blameless.
 - 10 See Todd (2011), Khoury (2014), and Tierney (2014) for discussion of related issues about manipulation and mitigation.
 - 11 McKenna (2008a) cites this as a reason why compatibilists should prefer a hard-line reply.
 - 12 For some recent soft-line replies, see Barnes (2013), Deery and Nahmias (2017), Mickelson, 2010 Schlosser (2015), and Waller (2014).
 - 13 This premise doesn’t appear in van Inwagen’s presentation of the argument. I’ve altered this aspect of the argument for the sake of simplicity. He appeals instead to the separate claims that “no one is even partly responsible for H” and “no one is even partly responsible L.” For a discussion of the logical relationship between the premise used here and van Inwagen’s two premises, see Warfield (1996).
 - 14 See Robinson (Forthcoming) for a response to Hermes along these lines.
 - 15 For one such argument, see Capes (2016).
 - 16 That’s not quite right. Kearns (2011, pp. 322–323) points out that if his criticism of 3 is correct, then we can be responsible for the laws of nature, in which case 5 is false.
 - 17 Ravizza (1994) and Fischer and Ravizza (1998, ch. 6) were the first to use overdetermination cases to challenge the principle. For some other discussions of 6, including attempts to provide counterexamples to it that don’t involve overdetermination, see Campbell (2006), Haji (2008, 2010), Shabo (2010b), and Widerker (2002).
 - 18 I also argue (Capes 2016) that this case can be used to refute the direct argument advanced by Warfield (1996). Stump and Fischer (2000) also respond to Warfield’s argument.
 - 19 See Widerker (2002) and Widerker and Schnall (2014).
 - 20 Shabo (2010a) endorses a similar principle. See Stump (2000, 2002) and Warfield (1996) for other attempts to provide alternative transfer of non-responsibility principles.
 - 21 See Schnall and Widerker (2012) for a response to McKenna. Yael Loewenstein (2016) develops McKenna’s view and responds to Schnall and Widerker.
 - 22 The following argument is similar to an argument in Widerker (2002).
 - 23 Thanks to Michael McKenna and Hannah Tierney for helpful comments.

Bibliography

- Barnes, E.C. (2013). Freedom, creativity, and manipulation. *Nous* 49 (3): 560–588. doi: 10.1111/nous.12043.
- Campbell, J.K. (2006). Farewell to direct source incompatibilism. *Acta Analytica* 21 (4): 36–49.

- Capes, J. (2013). Mitigating soft compatibilism. *Philosophy and Phenomenological Research* 87 (3): 640–663. doi: 10.1111/j.1933-1592.2012.00579.x.
- Capes, J. (2016). Incompatibilism and the transfer of non-responsibility. *Philosophical Studies* 173 (6): 1477–1495. doi: 10.1007/s11098-015-0559-1.
- Clarke, R. (2003). *Libertarian Accounts of Free Will*. New York: Oxford University Press.
- De Marco, G. (2016). Rescuing the zygote argument. *Philosophical Studies* 173 (6): 1621–1628. doi: 10.1007/s11098-015-0571-5.
- Deery, O. and Nahmias, E. (2017). Defeating manipulation arguments: interventionist causation and compatibilist sourcehood. *Philosophical Studies*. doi: 10.1007/s11098-016-0754-8.
- Fischer, J.M. (2011). The zygote argument remixed. *Analysis* 71 (2): 267–272.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control: An Essay on Moral Responsibility*. Cambridge: Cambridge University Press.
- Haji, I. (2008). Reflections on the incompatibilist's direct argument. *Erkenntnis* 68 (1): 1–19. doi: 10.1007/s10670-007-9056-z.
- Haji, I. (2010). On the direct argument for the incompatibility of determinism and moral responsibility. *Grazer Philosophische Studien* 80 (1): 111–130.
- Hermes, C. (2014). A counterexample to rule A. *Philosophia* 42 (2): 387–389. doi: 10.1007/s11406-013-9509-3.
- Kearns, S. (2012). Responsibility for necessities. *Philosophical Studies* 155 (2): 307–324. doi: 10.1007/s11098-010-9574-4.
- Kearns, S. (2012). Aborting the zygote argument. *Philosophical Studies* 160 (3): 379–389. doi: 10.1007/s11098-011-9724-3.
- Khoury, A. (2014). Manipulation and mitigation. *Philosophical Studies* 168 (1): 283–294. doi: 10.1007/s11098-013-0125-7.
- Loewenstein, Y. (2016). why the direct argument does not shift the burden of proof. *Journal of Philosophy* 113: 210–223.
- McKenna, M. (2001). Source incompatibilism, ultimacy, and the transfer of non-responsibility. *American Philosophical Quarterly* 38 (1): 37–51.
- McKenna, M. (2008a). A hard-line reply to Pereboom's four-case argument. *Philosophy and Phenomenological Research* 77 (1): 142–159. doi: 10.1111/j.1933-1592.2008.00179.x.
- McKenna, M. (2008b). Saying good-bye to the direct argument the right way. *Philosophical Review* 117 (3): 349–383. doi: 10.1215/00318108-2008-002.
- McKenna, M. (2014). Resisting the manipulation argument: a hard-liner takes it on the chin. *Philosophy and Phenomenological Research* 89: 467–484.
- Mele, A. (2005). A critique of Pereboom's 'four-case' argument for incompatibilism. *Analysis* 65: 75–80.
- Mele, A. (2006). *Free Will and Luck*. New York: Oxford University Press.
- Mele, A. (2008). Manipulation, compatibilism, and moral responsibility. *Journal of Ethics* 12 (3/4): 263–286. doi: 10.1007/s10892-008-9035-x.
- Mickelson, K. (2010). The soft-line solution to Pereboom's four-case argument. *Australasian Journal of Philosophy* 88 (4): 595–617. doi: 10.1080/00048400903382691.
- Mickelson, K. (2015). The zygote argument is invalid: now what? *Philosophical Studies* 172 (11): 2911–2929.
- Pereboom, D. (2001). *Living without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.
- Ravizza, M. (1994). Semicompatibilism and the transfer of non-responsibility. *Philosophical Studies* 75 (1): 61–93. doi: 10.1007/BF00989881.
- Robinson, M. (Forthcoming). Truthmakers, moral responsibility, and an alleged counterexample to rule A. *Erkenntnis* 81: 1333–1339. doi: 10.1007/s10670-015-9798-y.
- Sartorio, C. (2016). *Causation and Free Will*. New York: Oxford University Press.
- Schlosser, M.E. (2015). Manipulation and the zygote argument: another reply. *Journal of Ethics* 19 (1): 73–84. doi: 10.1007/s10892-014-9183-0.

- Schnall, I.M. and Widerker, D. (2012). The direct argument and the burden of proof. *Analysis* 72 (1): 25–36. doi: 10.1093/analys/anr134.
- Shabo, S. (2010a). The fate of the direct argument and the case for incompatibilism. *Philosophical Studies* 150 (3): 405–424. doi: 10.1007/s11098-009-9419-1.
- Shabo, S. (2010b). Against logical versions of the direct argument: a new counterexample. *American Philosophical Quarterly* 47 (3): 239–252.
- Stump, E. (2000). The direct argument for incompatibilism. *Philosophy & Phenomenological Research* 61 (2): 459–466. doi: 10.2307/2653663.
- Stump, E. (2002). Control and causal determinism. In: *The Contours of Agency: Essays in Honor of Harry Frankfurt* (ed. S. Buss and L. Overton), 33–60. Cambridge, MA: MIT Press.
- Stump, E. and Fischer, J.M. (2000). Transfer principles and moral responsibility. *Philosophical Perspectives* 14: 47–55. doi: 10.1111/0029-4624.34.s14.3.
- Tierney, H. (2013). A maneuver around the modified manipulation argument. *Philosophical Studies* 165 (3): 753–763. doi: 10.1007/s11098-012-9974-8.
- Tierney, H. (2014). Tackling it head on: how best to handle the modified manipulation argument. *Journal of Value Inquiry* 48 (4): 663–675. doi: 10.1007/s10790-014-9461-x.
- Todd, P. (2011). A new approach to manipulation arguments. *Philosophical Studies* 152 (1): 127–133. doi: 10.1007/s11098-009-9465-8.
- Todd, P. (2013). Defending (a modified version of) the zygote argument. *Philosophical Studies* 164 (1): 189–203. doi: 10.1007/s11098-011-9848-5.
- Todd, P. (2017). Manipulation arguments and the freedom to do otherwise. *Philosophy and Phenomenological Research* 92 (3). doi: 10.1111/phpr.12298.
- Tognazzini, N.A. (2014). The structure of a manipulation argument. *Ethics* 124 (2): 358–369. doi: 10.1086/673434.
- van Inwagen, P. (1980). The incompatibility of responsibility and determinism. *Bowling Green Studies in Applied Philosophy* 2: 30–37.
- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.
- Waller, R.R. (2014). The threat of effective intentions to moral responsibility in the zygote argument. *Philosophia* 42 (1): 209–222. doi: 10.1007/s11406-013-9476-8.
- Warfield, T.A. (1996). Determinism and moral responsibility are incompatible. *Philosophical Topics* 24 (2): 215–226. doi: 10.5840/philtopics199624221.
- Watson, G. (1987). Responsibility and the limits of evil. In: *Responsibility, Character, and the Emotions* (ed. F. Schoeman), 256–286. Cambridge: Cambridge University Press.
- Widerker, D. (2002). Farewell to the direct argument. *Journal of Philosophy* 99 (6): 316–324. doi: jphil200299632.
- Widerker, D. and Schnall, I.M. (2014). The direct argument for incompatibilism. In: *Libertarian Free Will* (ed. D. Palmer), 88–106. New York: Oxford University Press.

Freedom and Time Travel

RYAN WASSERMAN

Time travel raises a number of interesting questions relating to freedom. This chapter introduces one of the most famous paradoxes of time travel (Section 1) and outlines the standard response to that puzzle (Section 2). It then considers the implications of these issues for the doctrine of fatalism (Section 3), the analysis of ability (Section 4), and the extent of human freedom (Section 5).¹

1 The Grandfather Paradox

The Grandfather Paradox is one of the oldest and most familiar puzzles of time travel.² There are many different versions of the paradox, but the most famous example in the philosophical literature is due to David Lewis:

Tim ... detests his grandfather, whose success in the munitions trade built the family fortune that paid for Tim's time machine. Tim would like nothing so much as to kill Grandfather, but alas he is too late. Grandfather died in his bed in 1957, while Tim was a young boy. But when Tim has built his time machine and traveled to 1920, suddenly he realizes that he is not too late after all. He buys a rifle; he spends long hours in target practice; he shadows Grandfather to learn the route of his daily walk to the munitions works; he rents a room along the route; and there he lurks, one winter day in 1921, rifle loaded, hate in his heart, as Grandfather walks closer, closer. (1976, p. 149)

According to Lewis, there is a good reason for thinking that Tim *can* kill Grandfather in the envisioned scenario. After all:

He has what it takes. Conditions are perfect in every way: the best rifle money could buy, Grandfather an easy target only twenty yards away, not a breeze, door securely locked against intruders. Tim a good shot to begin with and now at the peak of training, and so on ... In short, Tim is as much able to kill Grandfather as anyone ever is to kill anyone. (1976, p. 149)³

However, there seems to be an equally good reason for thinking that Tim *cannot* kill Grandfather. After all, if Grandfather were killed, then Father would not be born. And if Father were not born, then Tim would not be born either. But, in that case, Tim would not be able to travel back in time

to kill Grandfather. Hence, time travel seems to imply the contradiction that Tim both can and cannot kill Grandfather.⁴ More generally, time travel would seem to allow for people to perform “self-defeating actions” – actions such that, were they performed, the relevant agent would not have existed. Since it is impossible for an agent to perform an action of this kind, it seems as if time travel must also be impossible. We will refer to this as *the self-defeating problem*.⁵

One of the most common replies to this problem is to invoke a “branching timeline” model of time travel, like the one depicted in *Back to the Future II*, *Star Trek*, and *Avengers Endgame*.⁶ On this view, when Tim travels back to the past, he ends up in an alternative version of 1921 where he is then able to kill Grandfather. This would make it the case that Tim is not born on the *alternative* timeline, but it would not change the fact that he was born on the *original* timeline. For this reason, Tim’s killing of Grandfather would not be a self-defeating act.

Unfortunately, the branching model of time travel is inconsistent with most of the traditional views about the nature of time. On the traditional way of thinking, there is a single 1921, and it either includes a killing of Grandfather or it does not. Either way, that fact is already settled and there is no way to change it now.⁷ If this view about the mutability of the past is correct, then there is a second reason for thinking that Tim cannot kill Grandfather. Here is how Lewis puts the point:

Grandfather lived, so to kill him would be to change the past. But the events of a past moment ... cannot change. Either the events of 1921 timelessly do include Tim’s killing of Grandfather, or else they timelessly don’t ... It is logically impossible that Tim should change the past by killing Grandfather in 1921. So Tim cannot kill Grandfather. (1976, p. 149)

This suggests a more general problem for the possibility of time travel. If time travel were possible, it would be possible to change the past. Since it is *not* possible to change the past, it is not possible to travel back in time. According to Lewis, this is the central problem raised by the grandfather paradox.⁸

2 The Standard Response

Generally speaking, there are three different responses to the grandfather paradox. First, one could accept the force of the argument and conclude that time travel is, in fact, impossible.⁹ Second, one could hold onto the possibility of time travel by challenging the claim that the past cannot be changed.¹⁰ Finally, one could resist the argument by insisting that – contrary to appearances – time travel would not put one in a position to change the past.¹¹

There are various ways of developing this last response, but the most familiar strategy is due to David Lewis (1976).¹²

Let’s begin by thinking about a very different kind of case.¹³ Suppose that Anna is a competent English speaker (and reader) who would like nothing more than to spend the afternoon with Jane Austen’s *Emma*. Unfortunately, someone has stolen the only available copy. Is it the case that Anna can still read *Emma*? On the one hand, it seems as if she *can*, since Anna has the general ability to read. On the other hand, it seems as if she *cannot*, since she lacks the opportunity to exercise that skill on the particular book in question. Lewis takes cases of this kind to show that ‘can’ is *context-sensitive*.¹⁴ On his view, to say that someone can do something is to say that her performing that action is compatible with all of the relevant facts. Since what is relevant will vary from context to context, uses of ‘can’ (and related terms like ‘able’) will express different things relative to different situations.

Here is one way of making this idea more precise. First, we analyze ‘can’-claims in terms of possibility. For example, we analyze (1) in terms of (2):

- (1) Anna can read *Emma*.
- (2) Possibly, Anna reads *Emma*.

Next, we understand possibility claims as generalizations about possible worlds. (2), for example, is understood as follows:

- (3) $\exists x$ (x is a possible world & Anna reads *Emma* in x)

Finally, we take context to provide a set of facts that limit the domain of quantification by restricting it to those worlds at which the relevant facts hold. This is what Angelika Kratzer (2012) calls the “modal base.” For example, in some contexts, we might wonder whether Anna has the intrinsic features required to read *Emma* (Are her eyes working? Does she understand English?) In those contexts, an utterance of (1) might express something like the following:

- (4) $\exists x$ (x is a possible world & Anna has the same intrinsic properties in x that she has in the actual world & Anna reads *Emma* in x)

This statement is obviously true – there are possible worlds in which Anna has her actual intrinsic properties and where she successfully reads *Emma* because her book is not stolen in those worlds. Suppose instead that we are curious whether Anna has the opportunity to read *Emma* in her current situation (for example, we might wonder whether she will be able to complete her reading assignment for class). In that case, (1) might express something like the following:

- (5) $\exists x$ (x is a possible world & Anna has the same intrinsic properties in x that she has in the actual world & Anna’s local environment in x is just like her environment in the actual world & Anna reads *Emma* in x)

Unlike (4), this statement is false – if Anna’s environment in another possible world is just like her environment in the actual world, then there won’t be any copies of *Emma* around. But, in that case, Anna won’t be reading the book.

Lewis takes this kind of context-sensitivity to be the driving force behind the grandfather paradox. Note, first, that the possibility of time travel would mean that there are some possible worlds in which Tim goes back in time and kills Grandfather. For example, there are presumably worlds in which Tim kills Grandfather, but where Grandfather miraculously rises from the grave (in order to sire Father, who will go on to sire Tim).¹⁵ Moreover, some of these worlds are ones in which Tim’s intrinsic properties and local environment are just like they are in the actual world (Tim has the same evil desires, the same make of gun, etc.). So, if we are in a context where the modal base is limited to these kinds of facts, the sentence ‘Tim can kill Grandfather’ will express a truth – his doing so would be compatible with all of the facts about his intrinsic properties and the local environment. Moreover, Lewis claims that these kinds of facts are the only ones that we normally hold fixed when assessing an agent’s abilities, so that Tim can kill Grandfather “By any ordinary standards of ability” (1976, p. 149).

Suppose, however, that we enrich the modal base by adding in the additional fact that *Grandfather will not be killed*. According to Lewis, this would not be an ordinary context, since we do not normally hold facts about the future fixed when assessing what an agent can and

cannot do. Still, there is nothing in principle to stop us from doing so and, in that case, the sentence ‘Tim can kill Grandfather’ would express the proposition that there is some world in which (a) Tim has the same intrinsic properties and local environment, (b) Tim kills Grandfather, and (c) *Grandfather is not killed*. Since there is no world in which (b) and (c) both hold, the relevant sentence would express a falsehood in the imagined context.¹⁶

However, the crucial point for Lewis is that there is *no* context in which ‘Tim can kill grandfather’ and ‘Tim cannot kill Grandfather’ *both* express truths, so the possibility of time travel does not generate a contradiction. Here is how Lewis summarizes these points:

Tim’s killing Grandfather that day in 1921 is compossible with a fairly rich set of facts: the facts about his rifle, his skill and training ... and so on ... But his killing Grandfather is not compossible with another, more inclusive set of facts. There is the simple fact that Grandfather was not killed ... You can reasonably choose the narrower delineation, and say that he can; or the wider delineation, and say that he can’t. But choose. What you mustn’t do is waver, say in the same breath that he both can and can’t, and then claim that this contradiction proves that time travel is impossible. (1976, pp. 150–151)

This is the standard response to the grandfather paradox.¹⁷

3 Time Travel and Fatalism

There is much to say in favor of the standard response. It is *general* in the sense that it provides a uniform solution to both the problem of self-defeat and the problem of past-alteration. It is *conservative* in the sense that it does not require a commitment to any controversial views about the nature of time, like the branching timeline model. And it is *independently motivated* since Lewis’s solution to the grandfather paradox can be applied to other problems as well. Consider, for example, the philosophical doctrine of *fatalism*. Fatalism is the view that whatever happens is unavoidable – that is, no one can do anything other than what they actually do.

Various arguments for fatalism have been put forward, but one of the oldest begins with the principle of bivalence. According to this principle, every meaningful statement is either true or false, including statements about the future. Consider, for example, the statement *that Ben will play with Legos tomorrow*. Suppose for the sake of argument that this statement is true. In that case, Ben cannot *avoid* playing with Legos tomorrow, since this would require him to make it *false* that he will play with Legos tomorrow and we have just assumed that this statement is true. Suppose instead that it is false that Ben will play with Legos tomorrow. In that case, Ben cannot avoid *not* playing with Legos tomorrow, since that would require him to make the relevant statement *true*. Once again, this would contradict our current assumption, since we have just taken it for granted that the statement is false. Putting these points together, it follows that Ben cannot avoid playing with Legos tomorrow or he cannot avoid *not* playing with Legos tomorrow. Moreover, there is nothing special about Ben and his actions, so we can generalize to every future event. In other words, everything is unavoidable. This is one way of putting the “logical” argument for fatalism.¹⁸

Most philosophers agree that this argument is unsound, but there is disagreement over where, exactly, it goes wrong.¹⁹ According to Lewis, the problem with this argument is very similar to the problem with the grandfather paradox. When the Fatalist invites us to temporarily assume (for the sake of argument) that Ben will play with Legos tomorrow, this has the effect of adding the statement *that Ben will play with Legos tomorrow* to the modal base (along

with, say, the facts about Ben's intrinsic properties and his local environment). As a result, an utterance of 'Ben can avoid playing with Legos tomorrow' will express something like the following (in the given context):

- (6) $\exists x$ (x is a possible world & Ben has the same intrinsic properties in x that he has in the actual world & Ben's local environment in x is just like his environment in the actual world & Ben plays with Legos tomorrow in x & Ben avoids playing with Legos tomorrow in x)

Since there are no worlds in which the last two conditions hold, the sentence 'Ben can avoid playing with Legos tomorrow' will express a falsehood, relative to that context. Similarly, if we temporarily assume (for the sake of argument) that Ben will *not* play with Legos tomorrow, then the sentence 'Ben can avoid *not* playing with Legos tomorrow' will express something like the following:

- (7) $\exists x$ (x is a possible world & Ben has the same intrinsic properties in x that he has in the actual world & Ben's local environment in x is just like his environment in the actual world & Ben does not play with his Legos tomorrow in x & Ben avoids not playing with his Legos tomorrow in x)

Once again, this statement is false. So, the sentence 'Ben can avoid not playing with Legos tomorrow' will express a falsehood, relative to the latter context. The problem, however, is that this context is very different from the previous one, since the Fatalist has changed the modal base. The same thing will be true when it comes to the conclusion of the argument (that Ben cannot avoid playing with Legos tomorrow or he cannot avoid *not* playing with Legos tomorrow). At this point, the Fatalist has discharged the assumption that Ben *will* play with Legos tomorrow, as well as the assumption that he *won't*. But, once we remove those statements from the modal base, the sentence 'Ben cannot avoid playing with Legos tomorrow or he cannot avoid *not* playing with Legos tomorrow' will express something like the following:

- (8) $\sim \exists x$ (x is a possible world & Ben has the same intrinsic properties in x that he has in the actual world & Ben's local environment in x is just like his environment in the actual world & Ben avoids playing with his Legos tomorrow in x) $\vee \sim \exists x$ (x is a possible world & Ben has the same intrinsic properties in x that he has in the actual world & Ben's local environment in x is just like his environment in the actual world & Ben avoids not playing with his Legos tomorrow in x)

This statement, however, is false – there are many worlds in which Ben (with his actual intrinsic properties and local environment) plays Legos and many in which he does not.

On this way of looking at things, the logical argument for fatalism is a bit like the following line of reasoning:

That is red.
That is green.
So, that is both red and green.

On the surface, this argument appears valid. But suppose that I am pointing at a red apple when I assert the first premise, a green apple when I assert the second premise, and a yellow banana when I assert the conclusion. In that case, the argument will obviously be invalid, since 'that' will pick out different objects on different occurrences.²⁰

4 Time Travel and Ability

Lewis's views on time travel and fatalism have proven very popular. However, they are not without controversy. The most famous objection to his account is due to Kadri Vihvelin (1996, 2011a, 2011b).

Vihvelin's objection is based on the following principle about abilities and counterfactuals:

The Counterfactual Principle (CP): Necessarily, if someone would fail to do something, no matter how hard or how many times she tried, then she cannot do it (i.e., she lacks the relevant ability). (2011a)²¹

Whether or not this principle is correct will depend, in part, on the modal base at issue. For example, we have already noted that there are contexts in which it would be correct to say that Anna can read *Emma*, even though there are no copies of the book for her to read. Clearly, that is not the kind of context Vihvelin has in mind, since (given the lack of books) Anna would invariably fail to read *Emma* no matter what she tried. Vihvelin is instead thinking of more restrictive contexts in which the modal base includes facts about the agent's environment, like the fact that there are no copies of *Emma* around.²²

Understood in this way, (CP) seems entirely plausible.²³ For one thing, many examples seem to conform to this pattern.

I can swim and ride a bicycle at least partly because it's true that I would succeed at both of these activities if I tried to do them ... But I cannot walk on water or run faster than the speed of light because if I tried to do these things, I would fail, no matter how often or how hard I tried. (1996, p. 318)

Moreover, this principle seems to underlie much of our practical reasoning about what we do:

Think about how you deliberate when you decide what to do. You choose among possible course of actions that you believe are your options ... But everyone agrees that a course of action is *an option for you* (something you have the *wide ability* to do) only if you would have some reasonable chance of doing it, if you tried, then and there, to do it. If the things you deliberate about are all things such that if you tried to do them, you would fail, you are wasting your time. (2011b)

The problem is that all of this is inconsistent with what Lewis wants to say about the case of Tim and Grandfather. Given Tim's actual causal history, we know that:

- (9) If Tim were to try to kill Grandfather, he would fail (no matter how hard or how many times he tried).

But then, given Vihvelin's principle, we can infer:

- (10) Tim cannot kill Grandfather.

More carefully, we can infer (10) from (9) in any context where (CP) expresses a truth. Since that principle arguably expresses a truth in many ordinary contexts, it will turn out that (10) is likewise true in many ordinary contexts. This, however, is inconsistent with what Lewis wants to say. On his view, (10) should only turn out true in "fatalistic" contexts where we

have built facts about the future into the modal base. If Vihvelin is correct, then Lewis is mistaken about this.

In order to appreciate Vihvelin's view, it is important to be clear about what she is and is not saying.

First, in disagreeing with Lewis about (10), she does reject the other parts of his theory. For example, it is perfectly consistent with everything Vihvelin says to think that 'can'-claims are restricted possibility claims that are sensitive to context in exactly the way envisioned by Lewis. The only disagreement between Lewis and Vihvelin is over which kinds of facts are normally held fixed when assessing abilities (and, in particular, whether or not those facts would be compatible with Tim killing Grandfather).

Second, in disagreeing with Lewis, Vihvelin is not saying that the grandfather paradox is sound. In fact, she explicitly rejects this argument.²⁴ Her view is that the very first premise of the argument is mistaken – time travel *is* possible (she says) but that does *not* mean that travelers are able to go back in time and kill their ancestors. Time travelers are simply limited in ways that ordinary individuals are not.

Third, in saying that Tim cannot kill Grandfather, Vihvelin is not saying that this action is metaphysically impossible. In fact, she explicitly says the opposite – for example, she says that there are worlds in which Tim travels back in time and kills Grandfather, only to see Grandfather miraculously rise from the grave (in order to sire Father, who will go on to sire Tim).²⁵ Vihvelin's point is that such a world would require very different laws of nature from our own, and is thus irrelevant to evaluating counterfactuals like (9).²⁶

Here it may be helpful to consider an analogy. Suppose, for the sake of argument, that there are some possible worlds in which I walk on water. Presumably, these are worlds with very different laws of nature than our own. For this reason, those worlds are irrelevant to the question of whether I *would* walk on water, were I to try. Intuitively, I would *fail* to walk on water no matter how hard or how many times I tried. So, (CP) implies (correctly) that I cannot walk on water. According to Vihvelin, the same thing is true for Tim. There might be *some* possible worlds in which Tim successfully kills Grandfather, but those worlds are not relevant to the question of whether or not Tim *can* kill Grandfather (not, at least, if we are working with a normal modal base).

Interestingly, Vihvelin has a very different view about other acts of past-alteration. Suppose, for example, that Grandfather was neither killed nor pinched during 1921. And suppose that Tim goes back in time to shoot Grandfather, but has a last second change of heart. Could Tim then choose to pinch Grandfather instead? Vihvelin thinks so. Presumably, Tim *will not* try to pinch Grandfather. But that does not mean he *cannot*. After all, there are many things we *can* do but *won't*. What would have happened if Tim had tried to pinch Grandfather? Presumably, he would (or at least might) have succeeded. The complete explanation for this is complicated,²⁷ but the short story is that pinching, unlike killing, would not be self-defeating. The only way for Tim to successfully kill Grandfather in the past would be if Grandfather were later resurrected from the dead (or if some other such miracle took place). That is because Grandfather's continued existence is required for Tim's existence in the past. The same thing is not true when it comes to pinching. If Tim would have pinched Grandfather, there would not have been any large-scale miracles or other surprising events – Grandfather would simply have gone on to sire Father, who would have sired Tim, who would have gone back in time and pinched Grandfather. For this reason, Vihvelin claims that time travelers can perform many actions that they do not actually perform in the past.

5 Time Travel and Freedom

A different worry for Lewis is suggested by Michael Rea (2015). Unlike Vihvelin, Rea argues that time travelers are not free with respect to *any* of their actions in the past. Indeed, Rea goes even further, arguing that the lack of freedom on the part of time travelers transfers to everyone else around them, so that “time travel destroys freedom on a global scale” (2015, p. 279).

Rea’s argument is based on two definitions, an observation, and a pair of principles. The definitions are as follows:

(D1) e_1 is a part of x ’s causal history at $t =_{\text{def}} e_1$ stands in the ancestral of the causal relation to an event e_2 that occurs at t and involves x as subject.

(D2) the causal history of x at $t =_{\text{def}}$ the sum of all events e such that, at t , e is a part of x ’s causal history.

To say that e_1 stands in the ancestral of the causal relation to e_2 is to say that there is a series of events, $a, b, c, \dots$ such that e_1 is a cause of a , which is cause of b , which is cause of c , which ... is a cause of e_2 . So, for example, my choice of college played a role in my studying philosophy, which played a role in my writing this paper. Hence, my choice of college is a part of my causal history, relative to the current time. And, of course, my causal history will go back even further, from my schooling to my birth and, ultimately, to the beginning of the universe.

Rea next observes that, for time travelers, events from one’s causal history can also be within one’s temporal future. This might sound paradoxical at first, but it is exactly what we should expect in the case of backward time travel. After all, in that kind of case, the traveler’s departure is a cause of his arrival, even though the arrival precedes the departure in time.²⁸

Rea argues that this observation leads to the conclusion that time travelers are limited in surprising ways. His argument is based on the principle that “no one is able to do anything precluded by her own causal history” (2015, p. 272). More carefully:

The Causal History Principle (CHP): For any agent S , action y , and time t , if something about S ’s causal history (relative to t) would have been different, were S to do y at t , then S is not able to do y at t .

(CHP) is closely related to the famous “Fixity of the Past” principle that is featured in the traditional argument for the incompatibility of freedom and determinism.²⁹ According to that principle, the past is “fixed” in the sense that no one can do anything such that, were they to do it, some fact about the past would have to have been different from how it actually is.³⁰ Rea’s suggestion is that the same thing is true when it comes to an agent’s *causal* past. This is what leads to trouble for time travelers. Consider, again, the case of Tim. Tim’s refraining from killing Grandfather (at t_2 , let’s say) is a part of his causal history at t_1 (as he takes aim at Grandfather). So, given (CHP), Tim is unable to kill Grandfather, even though he has a loaded gun, a clear shot, etc.

The final piece of Rea’s argument is the claim that “Everybody shares the same causal history up to some particular time in the fairly recent past” (2015, p. 276). More carefully:

Local Holism (LH): For any time t in evolutionary history on earth, all person-stages p that exist at t belong to a system s of causally related events of which the following condition holds: there is some duration d such that, for any event e in s , if the duration between e ’s occurrence and t is greater than or equal to d , then every event in e ’s causal history is also in p ’s causal history. (2015, pp. 276–277)

The statement of this principle is complicated, but the basic idea is simple. If I raise my hand now, that will have a subtle impact on all of the nearby air molecules, which will eventually have an impact on all of the air molecules in the next room, which will then have an impact – however small – on the particles that make up my son’s body. Hence, the raising of my arm will soon be a part of my son’s causal history. The same thing will be true for his recent movements and my causal history. More generally, if an event stands in the ancestral of the causal relation to an event involving one person, it will (sooner or later) stand in the same relation to an event involving any other arbitrarily selected individual.

Let us now apply this to the case of Tim when he travels to the past. Suppose that Tim does not shoot his grandfather, and instead chooses to pinch him. That pinch will have an immediate effect on Grandfather’s body – perhaps brushing off a few skin cells or slightly altering the position of some other particles. Given (LH), these effects will have their own effects on the people around Grandfather, which will then have effects on other people, which will – ultimately – have an effect (however small) on Tim himself. For example, we can imagine that Tim’s pinching of Grandfather is a cause of a cause of a cause ... of a particular particle being in a specific place on Tim’s right hand. But that means that, when Tim is deciding whether or not to pinch Grandfather, the future pinching is already a part of his causal past – Tim already bears the mark (however small) of the future decision. But then, given (CHP), it follows that Tim cannot refrain from pinching Grandfather. This is a surprising result.

And things get worse. Suppose that, a few minutes after Tim pinches Grandfather, a person at the nearby café takes a sip of coffee. Given (LH), this event will have an immediate effect on the nearby air molecules, which will then have an impact on the nearby people, which will eventually have an impact (however small) on Tim in the future. For example, we can imagine that the position of a specific particle on Tim’s ear stands in the ancestral of the effect relation to the person’s sipping coffee in 1921. But then, once Tim has traveled into the past, the position of the relevant particle will have an effect (however small) on other nearby people, including the person at the café who has just ordered her coffee. But that means that the person’s future sipping of the coffee is already a part of her causal past – for example, it might be a cause of a cause of a cause ... of a specific particle being in a particular place on her nose. But then, given (CHP), the individual cannot refrain from sipping her coffee.

The general point is that, given (LH), backward time travel creates a multitude of subtle causal loops, which then constrain the abilities of everyone around the time traveler in surprising ways. If we add to this the familiar idea that freedom requires the ability to do otherwise,³¹ it would follow that backward time travel destroys freedom – not just for the time traveler, but for everyone around them. Indeed, if the traveler goes back far enough into the past, then, given the truth of (LH), they will destroy freedom for everyone.³²

6 Conclusion

As this brief survey indicates, the possibility of time travel raises a number of interesting questions about fatalism, ability, and freedom. It also intersects with more general debates about modality, counterfactuals, and causation. For all of these reasons, time travel has been a topic of much interest to philosophers, and will continue to be so for the foreseeable future.

Notes

- 1 For an alternative introduction to these topics, see Tognazzini (2016).
- 2 For the history of this paradox, see Wasserman (2018, Chapter 1 and Chapter 3, Section 1).
- 3 Lewis gives another argument for the same conclusion. For discussion of this argument, see Wasserman (2017).
- 4 We will be focusing on the case of backward time travel (rather than forward time travel), so we will leave the “backward” implicit from this point onward.
- 5 For a more careful characterization of self-defeating acts, see Law and Wasserman (2022).
- 6 The branching timeline view is closely related to the “multiverse” model of time travel that has been suggested by physicists like David Deutsch (1991). For more details on the branching model, see Wasserman (2018, Chapter 3, Section 3).
- 7 See, for example, Cook (1982, p. 49), Horwich (1975, p. 436), and Hoppers (1967, p. 177).
- 8 For further discussion of this “past-alteration” problem, see Wasserman (2018, Chapter 3, Section 2). For more on the relationship between self-defeat and past-alteration, see Sider (2002), Vihvelin (2011b), and Wasserman (2018, Chapter 4, Section 2).
- 9 This seems to be the conclusion favored by Grey (1999, Section 10).
- 10 This is one way of understanding the branching timeline model from the previous section. See also Goddu (2003), Loss (2015), and van Inwagen (2010).
- 11 See Vihvelin (1996).
- 12 Similar suggestions have been made by Fitzgerald (1974), Horwich (1975, pp. 435–437), and Thom (1975), among others.
- 13 See Austin (1956).
- 14 An alternative would be to say that ‘can’ and related terms like ‘ability’ are ambiguous. For example, we could say that Anna has the “general” ability to read *Emma*, but lacks the “specific” ability to do so in her current situation. Hence, there is one sense in which she can, and a different sense in which she cannot. (See Honoré (1964), Mele (2002), and van Inwagen (1983, p. 13)). In fact, in his original article, Lewis himself claims that ‘can’ is equivocal, with multiple distinct meanings (1976, p. 150). However, his considered view is that the term is unambiguous but relative. See Lewis (1979, pp. 354–355).
- 15 For other examples of this kind, see Carroll (2016).
- 16 Note that this move involves more than a simple distinction between “general” and “specific” abilities (see n.14). In order to get a context in which it is false that Tim can kill Grandfather, we cannot simply add in details about the local environment (as in the case of Anna and *Emma*); rather, we must build in facts about the future, like the fact that Grandfather will not be killed.
- 17 This response has been endorsed by many, including Hanley (1997, pp. 209–210), Sider (1997, p. 143), and Smith (1997, p. 366).
- 18 This kind of argument is suggested by Aristotle in the ninth chapter of *De Interpretatione*. For a more recent articulation and defense of this argument, see Taylor (1962).
- 19 For some different diagnoses, see Conee (in Conee and Sider 2005, pp. 29–30), Prior (1967, pp. 128–129), and Purtill (1988).
- 20 For more on the relationship between contextualism, fatalism, and freedom see Carroll (2010).
- 21 (CP) is reminiscent of the conditional analysis of ability, but that analysis is not forced upon the proponent of (CP). On the conditional analysis, see Ayer (1954).
- 22 In Vihvelin’s terminology, (CP) only applies to “wide” abilities, which take extrinsic, environmental factors into account. She contrasts these with “narrow” abilities that “you have in virtue of what’s beneath your skin” (2011a). This parallels our earlier distinction between “specific” and “general” abilities (see n.14).
- 23 For discussion of (CP), see Sider (2002), Vranas (2010), and Wasserman (2018, Chapter 4, Section 2).
- 24 Vihvelin (1996, p. 323).

- 25 Vihvelin (1996, p. 317).
- 26 For a more complete explanation of this point (including a more general discussion of how laws of nature figure into the truth-conditions for counterfactuals) see Vihvelin (2011b) and Wasserman (2018, Chapter 4, Sections 2.1 and 2.3).
- 27 See Law and Wasserman (2022), Vihvelin (2011b), and Wasserman (2018, Chapter 4, Sections 2.1 and 2.3).
- 28 Indeed, this kind of mismatch is often taken to be *definitive* of backward time travel. See, Lewis (1976, p. 146).
- 29 Here I have in mind the kind of argument given by van Inwagen (1983).
- 30 See Fischer (1994, p. 78).
- 31 See van Inwagen (1999).
- 32 For further discussion of Rea's argument, see Wasserman (unpublished).

Bibliography

- Austin, J.L. (1956). Ifs and cans. *Proceedings of the British Academy* 42: 107–132.
- Ayer, A.J. (1954). Freedom and necessity. In: *Philosophical Essays* (ed. A.J. Ayer), 271–284. New York: St. Martin's Press.
- Carrol, J.W. (2010). Context, conditionals, fatalism, time travel, and freedom. In: *Time and Identity* (ed. J.K. Campbell, M. O'Rourke and H.S. Silverstein), 79–94. Cambridge: MIT Press.
- Carrol, J.W. (2016). Ways to commit autoinfanticide. *Journal of the American Philosophical Association* 2: 180–191.
- Conee, E. and Sider, T. (2005). *Riddles of Existence*. Oxford: Oxford University Press.
- Cook, M. (1982). Tips for time travel. In: *Philosophers Look at Science Fiction* (ed. N.D. Smith). Chicago, IL: Nelson-Hall: 47–56.
- Deutsch, D. (1991). Quantum mechanics near closed timelike lines. *Physical Review D* 44: 3197–3217.
- Fischer, J.M. (1994). *The Metaphysics of Free Will*. Oxford: Blackwell.
- Fitzgerald, P. (1974). On retrocausality. *Philosophia* 4: 513–551.
- Goddu, G.C. (2003). Time travel and changing the past: (or how to kill yourself and live to tell the tale). *Ratio* 16: 16–32.
- Grey, W. (1999). Troubles with time travel. *Philosophy* 74: 55–70.
- Hanley, R. (1997). *The Metaphysics of Star Trek*. New York: Basic Books.
- Honoré, A.M. (1964). Can and can't. *Mind* 73: 463–479.
- Horwich, P. (1975). On some alleged paradoxes of time travel. *Journal of Philosophy* 72: 432–444.
- Hospers, J. (1967). *An Introduction to Philosophical Analysis*. London: Routledge & K. Paul.
- Kratzer, A. (2012). *Modals and Conditionals*. Oxford: Oxford University Press.
- Law, A. and Wasserman, R. (2022). Lessons from Grandfather. *Philosophies* 7:11.
- Lewis, D. (1976). The paradoxes of time travel. *American Philosophical Quarterly* 13: 145–152.
- Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic* 8: 339–359.
- Loss, R. (2015). How to change the past in one-dimensional time. *Pacific Philosophical Quarterly* 96: 1–11.
- Mele, A. (2002). Agents' abilities. *Noûs* 37: 447–470.
- Prior, A.N. (1967). *Past, Present and Future*. Oxford: Oxford University Press.
- Purtil, R. (1988). Fatalism and the omnitemporality of truth. *Faith and Philosophy* 5: 185–192.
- Rea, M. (2015). Time travelers are not free. *Journal of Philosophy* 112: 266–279.
- Sider, T. (1997). A new grandfather paradox? *Philosophy and Phenomenological Research* 57: 139–144.
- Sider, T. (2002). Time travel, coincidences and counterfactuals. *Philosophical Studies* 110: 115–138.
- Smith, N.J.J. (1997). Bananas enough for time travel? *British Journal for the Philosophy of Science* 48: 363–389.
- Taylor, R. (1962). Fatalism. *Philosophical Review* 71: 56–66.

- Thom, P. (1975). Time travel and non-fatal suicide. *Philosophical Studies* 27: 211–216.
- Tognazzini, N. (2016). Free will and time travel. In: *The Routledge Companion to Free Will* (ed. M. Griffith, N. Levy and K. Timpe), 680–690. New York: Routledge.
- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Oxford University Press.
- van Inwagen, P. (1999). Moral responsibility, determinism, and the ability to do otherwise. *The Journal of Ethics* 3: 341–350.
- van Inwagen, P. (2010). Changing the past. *Oxford Studies in Metaphysics* 5: 3–40.
- Vihvelin, K. (1996). What time travelers cannot do. *Philosophical Studies* 81: 315–330.
- Vihvelin, K. (2011a). Ability, “can”, and counterfactuals. Vihvelin.com, accessed on 11th October 2021, <http://vihvelin.typepad.com>.
- Vihvelin, K. (2011b). Counterfactuals, indicatives and what time travelers can’t do. Vihvelin.com, accessed on 11th October 2021, <http://vihvelin.typepad.com>
- Vranas, P. (2010). What time travelers may be able to do. *Philosophical Studies* 150: 115–121.
- Wasserman, R. (2017). Time travel, ability, and arguments by analogy. *Thought* 6: 17–23.
- Wasserman, R. (2018). *Paradoxes of Time Travel*. Oxford: Oxford University Press.
- Wasserman, R. unpublished. Time travel, freedom, and incompatibilism.

Divine Freedom

BRIAN LEFTOW

The three Abrahamic monotheisms' main streams agree on a great deal about God. They agree e.g. that God is immaterial, all-knowing, all-powerful, all-wise, eternal, the maker and sustainer of all concrete things He does not contain, and in providential control of history. I now take up what picture of divine freedom best fits this shared Abrahamic concept of God. I first suggest that God so viewed is uniquely free of external constraint. I next consider attributes theories of freedom take to be constitutive of being free in other senses, and argue that God has them in excelsis. I consider two views of the extent of His ability to do otherwise, and an argument that due to His goodness, He is not free to actualize any possible world.

1 External Constraint

By “constraint” I mean a state of affairs that impedes or prevents one’s trying to do or doing something. A constraint is external if it does not come from one’s nature, one’s parts, or oneself as a whole. Thus external constraints on God might be in principle be logic and mathematics, the laws of nature, particular matters of natural fact, the free choices or actions of His creatures, and/or their consequences.

2 Logic and Mathematics

Pace Descartes, even an omnipotent being cannot violate logic or mathematics. I suggest elsewhere (with e.g. Aquinas and Leibniz) that facts of logic and mathematics are not external to God (Leftow 2012). Rather, they arise out of God’s own nature. Even if I am wrong about that, these cannot count as a constraint. Logic and mathematics mark lines where possibility runs out: it is not possible that $2 + 2 \neq 4$. Powers can only effect what is possible. “Can do the logically impossible” seems contradictory – if someone can bring it about, it isn’t impossible. Thus there cannot be a power to bring it about that $2 + 2 \neq 4$ unless this is possible. Descartes’s notorious claim that “it was free and indifferent for God to make it not be true ... that contradictories could not be true together” (Kenny 1970, p. 150) is perhaps most

charitably taken as implying that this is possible. But it isn't. It is no constraint not to have a power there cannot be.

3 Natural Laws

What to say about God and natural law depends on what natural laws are. Suppose first that they are broadly Humean – contingent generalizations about all particular matters of physical fact.¹ For Humeans, laws don't "govern."² They result, when all the facts are in. It's standard Abrahamic theism that God controls what the particular physical facts are. It's up to Him how He runs nature. So on this account, laws do not constrain Him.

Suppose that natural laws are contingent, but "govern." For instance, suppose that they involve a contingent nomic-necessitation relation between having one property and having another, which obtains independent of what the particular physical facts are. Then God's omnipotence gives Him power over what the laws are. For as standardly understood, being omnipotent is (at a first pass) having the power to actualize any contingent state of affairs not dependent on others' libertarian free will.³ So He could have nomically linked any properties which could be so linked. Our intuitions that other laws are possible suggest that others can indeed be. So again, it is up to God what the laws are. If they are any constraint, it is self-imposed, for His own reasons.

Finally, suppose that the laws are metaphysically necessary. Suppose, say, that it is metaphysically necessary that in certain sorts of circumstance, iron rusts. At least within Christianity, it is a widespread view that necessary truths' truthmakers lie somehow in God – in His nature (Aquinas, Scotus, Leibniz), or some determination of His will (Descartes and perhaps Anselm).⁴ Having one's nature is not an external constraint if the nature is not externally imposed. God's is not. So on this view, if the laws constrain God, that constraint either is not external or is self-imposed (Anselm and Descartes again). Further, there remains a sense in which the laws are up to Him, even if necessities are due to His nature. For it is up to Him whether there is any iron. So it is up to Him whether He has to deal in practice with the necessity that iron rusts.

Further, however we think of natural law, it is standard Abrahamic theism that God is non-physical. So they do not govern His being. Again, whatever laws' status, it is standard Abrahamic theism that God leaves room in nature to do miracles. This might mean that the laws contain clauses we would not have expected, or that they are really (say) truths about what natural powers naturally do, and God can give creatures other powers on occasion.

Logic and mathematics constrain no-one. Humean laws constrain no-one. Many particular facts that determine them constrain finite agents, but do not constrain God. On other views of laws, only for God are natural laws not external constraints.

4 Particular Fact: The Past

I now take up particular facts that might constrain God. The main category one might point to here is the past. Medieval philosophical theologians were unanimous that the past is unalterable.⁵ They thought this a metaphysical necessity.⁶ So to them, at least, it was no more an external constraint than arithmetic or the law about iron rusting.

Recently Hud Hudson has proposed a model that gives God power over the past (Hudson 2014). Suppose that our time is a 4D eternalist block – a spacetime all of which is always there. Suppose too God lives in a flowing hypertime – a time like the one we think we live in.

Then God might make the 4D block at hypertime t , and replace part of it with something else at hypertime $t + 1$. If so, for anything after that part, God has changed the past. For some P , hyperfirst it was the case that P , and hyperthen it was the case that not- P . If Hudson is right, the past is indeed alterable, and God is not constrained at all by its content. Of course, it is one thing to have the power to do this, and another to have any reason to do it.

On less outré pictures of time, it's hard to see how the past could be changed. Suppose, for instance, presentism – that only the present is real. For presentists, only what presently exists falls under the absolutely unrestricted existential quantifier. Equivalently, if one absolutely de-tenses “exists,” for presentists, only what is present exists. Thus for presentists, the truthmakers for truths about the past must lie in the present. For some, they are presently existing Platonic abstract entities (e.g. Bourne 2006). For others, it may be that present concrete things have attributes like being such that the Germans lost WW2 (Merricks 2007, p. 133ff.). I do not see how anyone could act on a Platonic entity to change it. Nor do I see how anyone could act on something presently existing to make it not the case that the Germans lost WW2. But on presentism, that is what it would take to change the past.

Now suppose that not just the present, but also the past exists. Then for a temporal being to change the past would take some sort of backward causation. While friends of time-travel disagree, most consider backward causation impossible. If God is atemporal, all of time is spread before Him (Boethius 1973, V, 6). It's all real, from His standpoint; also, from His standpoint, He is creating it all. I can't see, myself, how it would “look” to such a God to have the past change. As in Hudson, it seems that there would have to be some sort of hyperchange in what God “sees.” A change in God's seeing is a change in God. An atemporal God can't change. So it seems that an atemporal God couldn't change an existing past either.

Thus it seems that God could not change the past, whether it exists or not and whether He is atemporal or not. But it does not follow that the past imposes an external constraint on Him. The Abrahamic religions agree that if God is atemporal, all of time is His creation. If it is, the past's being past is no constraint on Him at all. For the standpoint from which He acts is causally prior to anything's being past. From His own standpoint, He is always timelessly making the past what it is. He is passing up His chance to have it otherwise, but it is no more a constraint on Him to freely pass it up than it is a constraint on us to freely have eggs rather than cereal for breakfast. On the other hand, if God is by nature temporal, plausibly time is not an external constraint on Him. For plausibly, time is a consequence of God's nature. By nature God is temporal. That is, by nature, He persists. Thus one phase of His life succeeds another. This is why there is such a thing as succession. On such a view, God's nature imposes the fixity of the past. So any constraint here is not external.

The fixity of the past constrains anyone but God. Only the divine attributes make it not an external constraint.

5 Creaturely Freedom

God must (as it were) work around creatures' freedom only if it enables them to do things God does not determine them to do. If it does, they have it because He gave it. He gave it knowing all its possibilities – everything they might do with it. Further, these are possibilities only because He permits them. So any obstacle we pose for God, He permits in advance, for His own reasons. It is thus in a way self-imposed, though by means of our freedom.

God always knew all the possibilities. What happens is always something He has already considered and permitted. So unanticipated circumstances – things He hadn't thought of – could

not force Him to act. Further, He can always destroy anything. So He can overrule any other concrete thing, at least by eliminating it.⁷ This creates a thicker sense in which any obstacle we pose is only by His permission. Suppose that I put Him in a dilemma. If He lets me live, I will freely do evil He dislikes. But He equally dislikes killing me or interfering with my freedom or circumstances to prevent this. Then either way, God must choose an alternative He would rather not choose. That's why it's a dilemma. But God accepted the possibility of the dilemma in advance. Further, He could have avoided it. So the constraints of the dilemma are self-imposed.

I have examined the possible sources of external constraint on God's action. It should be clear that God is uniquely free of this.

6 Theories of Freedom

Freedom from external constraint is one form of freedom of action. One way to a full account of God's freedom would be to define other sorts of freedom and show that God satisfies the definitions. I now instead show how God has features important in many accounts of freedom. It will be clear that having these as He does suffices for Him to be free in many reasonable senses.

Libertarian accounts of freedom often appeal to being an ultimate source of one's action, ability to do otherwise, and control of one's actions (e.g. Kane 1996). Compatibilist accounts speak of a different sort of control (Fischer and Ravizza 1998) and of a "mesh" between the acts one does and one's beliefs, desires and values (Wolf 1990), or of acting on desires one wants to have (Frankfurt 1971), or of being responsive to reasons (Fischer and Ravizza 1998).

7 Ultimate Sourcehood

Intuitively, I act freely only if my action really comes from me. Accounts of freedom differ in how they accommodate that intuition. Some do so in light of another intuition, that if causes outside me fully cause my action, it originates outside me, in those causes. If so, it does not in the fullest sense come from me. Thus some libertarians claim that one acts freely only if there are no such causes.

I suggest that

God is an ultimate source of action = df. God can do an act which has no
causal conditions wholly outside Him, from
which He was fully able to refrain.

The first clause provides that nothing outside God determines this action. The second provides that nothing inside God determines it either. Instead, it comes from Him. Even if He has "compelling" reason to do the act, He, not His reasons, has the final say. For the Abrahamic religions, God is in this sense the ultimate source of His act of creating. Before He creates, there is nothing outside Him to make Him do so. So He satisfies the first clause. The Abrahamic religions hold that God could have refrained from creating. So on their view, at least, He satisfies the second too. It is standard Abrahamic theology that God is the ultimate source of all His actions. For it is standard to hold that God is impassible – that nothing can else cause Him to do anything, though other things can give Him reasons to act.

Freedom of action is one thing. Freedom of will is another. One has it, obviously, only if one has a will. So I now ask whether God has the powers we speak of collectively as a will – to decide, choose, and form and maintain intentions.

8 Does God Have a Will?

The usual view is that He does. But Linda Zagzebski suggests that God has no will and makes no decisions. Rather, God just has motivations – emotions – that collectively point to acting one way rather than another, and these dictate what He does (Zagzebski 2004, pp. 217–219, 293–295). Zagzebski's argument has two premises, that there is no needless mental machinery in God, and that God has no need of a will. Her case for the second is this. A will's "jobs" are (says she)

- (I) to move us from indecision to decision (*ibid.*, 294),
- (II) to distinguish voluntary from involuntary actions (the voluntary are those one wills to do) (*ibid.*, 291–2),
- (III) to enable us to overrule causal forces threatening to determine our actions and so safeguard libertarian freedom (*ibid.*), and
- (IV) to resist desire and emotion (*ibid.*, 293).

But (says Zagzebski) *contra* (i), a perfect being is never undecided, and so need not pass from that to decision (*ibid.*, 294). *Contra* (ii), a perfect being does not act involuntarily, and so there is no need to distinguish voluntary from involuntary (*ibid.*, 293). *Contra* (iii), nothing outside God can determine His actions and "there is no question of God's motives being put upon Him or involuntarily suffered" (*ibid.*, 293). *Contra* (iv), a perfect being's emotions and desires are not the sort of thing He needs to rise above (*ibid.*).

I think a divine will would have other jobs than (i)–(iv). God could face "Buridan's Ass" choices. He could have precisely equal motivations for mutually exclusive acts. He might e.g. find two mutually exclusive actions equally good and each better than all alternatives save the other action. When Elizabeth prays for a child, for instance, He might decide to give her either John the Baptist or someone else precisely as worth having. When God decides what universe(s) to make, He might need a will to deal with a different sort of impasse. For plausibly, for every universe, a better is possible. Plausibly, there is no best possible universe. So for every universe, He has reason to create it, and also reason to create something better instead. Making a multiverse consisting of all universes above some cutoff in goodness is not a unique best option either. For different cutoffs are possible, none obviously best. Moreover, God still has to decide how many qualitative duplicates of the multiverse to make. For each, more would be better.

It would be a defect in the divine design were there no way to make such choices. So if God is a perfect being, there would be a way. Perhaps He would actively decide, by one of the powers we collect under the rubric "will." If not, the transition from merely having motivation to acting would "just happen" in Him, and its happening would be what we (misleadingly) call His deciding. If it "just happened" by the uncaused forming of an intention, that too would involve a will. So if God has no will, seemingly it would have to "just happen" by God's motivations indeterministically causing His action. (That is, they would cause it, but what they cause might have differed.) On this picture, nothing in God determines what His motives cause. It "just happens" to be this or that. If a perfect being is never in a state of indecision, then at any time, this "just happening" either is happening or has happened.

The active decision picture seems preferable. It avoids the odd thought that something God causes is nonetheless ultimately down to sheer chance, the chance that God's motives caused it, not another thing. It also gives God more control over what He does, and that is a good thing. On the "just happen" picture, it's hard to see how God could be morally responsible for His "decisions." They "just happen" to Him. They are internal accidents, not actions. So He cannot rightly be praised or blame for them. If they produce an effect, He cannot be praised or blamed for that, either. If I accidentally slip and so knock you down, I am not to blame. The case may be the same if the accident is internal, a "decision." If it is not, that is because it takes time to implement my decision, and during that period I may be able to countermand it – I see my fist travel toward you and have a split-second to realize what I'm doing and perhaps draw back. But if God simply decides "let x occur," x occurs at once. So He could not be praised or blamed for x . Perhaps God is praiseworthy for the state of His motivations. If He is, and these make His only live options A and B, He is praiseworthy for having only these options. But even if He is, He does not deserve praise for doing (say) A.

My next consideration also falls broadly under the heading of self-control. If God has no will, just two things in God might keep His strongest single first-order motivation from issuing in action. One would be a sort of vector sum of individually weaker first-order motivations. The other would be a second-order emotion, an emotion about emotions. God as Zagzebski depicts Him could in principle (say) feel shock at His strongest single motivation, and have the shock overrule it.

As what Zagzebski calls motivations are just emotions construed a particular way, on this picture God is literally the plaything of His emotions. He does not control what He (first-order) feels at the time He feels it any more than we do, though if He is temporal He might at earlier times have had indirect control over what He would feel now. Any control in the transition from first-order feeling to action is just some emotion(s) overcoming others. His emotions just happen to Him. So too does it just happen to Him that some are strong enough and rightly directed to overcome others. "Control" by second-order emotion just gives us one more emotion of which God is the plaything. Nor can God affect, nor then control, what His overall set of emotions lead Him to do, there being nothing outside the set to do so. There is literally nothing He can do about His emotions as He has them, because all doing occurs only after all emotions have weighed in.

Further, the controlling emotion – shock, in my example – either is or is not subject to further emotional approval before it initiates or blocks a divine action. If it is, an infinite regress of approvals of approvals begins. It would be far more economical to have a power to render a verdict on all God's emotions in a single act of decision. If it is not, God lacks even an emotion-based version of self-control over shock (or whatever). This seems an imperfection. Without adequate self-control, do we really have voluntary action?

If God has a will, we avoid all this. It gives Him clear responsibility for Buridan-case and infinite-hierarchy decisions. It avoids chance in God. It makes Him praiseworthy for what He does. Even if He never does resist the vector sum of all His motivations, He has the capacity. He "goes with the flow" only if He sees (as perhaps He always would) no need to resist it. He has self-control, in the sense that He can overrule all His emotions, even if He never has to use it. In the same way, a vigilant mother watching children scamper over a playground has them under control as long as she remains able to corral them if she wishes, even if she never wishes.

Perhaps God's emotions always initiate His actions. But perhaps not. God made n units of matter, not $n + 1$. Why n ? It's hard to see what could make n more attractive. If nothing could, God could only have more "feeling" for n as a pure matter of taste. He might just like

n better just because it is n, as one might just like chocolate better than other flavors just because it tastes like chocolate. But there are infinitely many such apparently arbitrary facts about the universe. It strains credulity to suppose that God had a unique strongest feeling about each of them. It seems more plausible that God just picked: in deciding between n and its alternatives, God made a Buridan's Ass choice. If this is how it is, God's reasons, motives and preferences take Him partway to a decision, but they do not determine it. The rest comes from somewhere. The best idea of where is that God has, and uses, a will. If the universe has some "just-pick" features, God's creative/providential decision as a whole is a just-pick. This is so even if the universe includes features for which He had determining reason. Those reasons just limited the field from which He picked.

If God's reasons, motivations etc. do not single out one action, or they do but He can refrain, He has a choice. His will has a role, if He has one. He has greater control of His actions if He does. Suppose on the other hand that God's motivations do uniquely point to one act – say, He finds Himself liking universes with red dwarfs much more than He likes any without them or there being no universe. On Zagzebski's model, God's emotions form a vector sum and yield a red dwarf universe automatically. If God has a will, it goes this way. God has various-strength attitudes to individual states of affairs. He strongly likes the idea that P, weakly dislikes the idea that Q, etc. Such attitudes are the basis of preferences.⁸ God's decisions express His preferences. God might (so to speak) run a check at two points. He might review His likes and dislikes in forming His preferences. He also might review His preferences before deciding, in each case assuring Himself that they conform to His standards. As we've seen, such checks could be on all likes and preferences and yet maximally economical only if they involve a will. Perhaps a perfect being has no likes He would not let figure in a preference, nor preferences He should not respect. So perhaps He would have no need to run a check, nor then for a power to do so in the most economical way. But still, He has greater control if He has the power to do it. And there are other reasons for God to have a will, as we've seen. So it seems best to say that He has one. For reasons we have already seen, it will be able to act entirely without prior external causal conditions. This provides a significant degree of freedom.

9 Control

Nothing makes God, gives Him His nature, compels or coerces Him. His nature guarantees that He knows everything external. Beyond that, external influences (e.g. creatures' libertarian-free actions, if any) constrain Him only if He permits. Omniscience guarantees that God fully grasps all the facts, and the reason-giving force of any fact. Perfect wisdom or rationality guarantees that He acts on His reasons. A reasonable thought, in fact, is that God always acts on all reasons He has (Pruss 2013). If so, His choices are maximally "reasons-responsive." This would guarantee a perfect mesh between His actions and His beliefs, values, etc. Finally, omnipotence presumably includes the power to eliminate desires. So another plausible thought is that God has only desires He approves of ("wants to have"), if He has desires.⁹ If so, these are the only ones He acts on. Even if not, omnipotence presumably guarantees enough self-control to guarantee that He acts only on desires of which He approves. For an omnipotent being can rid Himself of desires just by deciding to do so.

God has complete control of His actions and will. Nothing can cause, interfere with or modify His forming or executing His intentions, and He controls all external conditions under which He wills and acts. Thus He has greater control of His actions and will than any created agent could.¹⁰ I now discuss God's ability to do otherwise.

10 Ability to Do Otherwise

One can do otherwise, in the sense relevant to libertarian freedom, just if at some point in one's life one has more than one open alternative. One has a chance to take it: all conditions allow more than one choice. One has all needed powers. And nothing prevents one's using them in more than one way (which is really just part of having the chance). Theists generally have held that God could have chosen differently than He actually has. In Islam, divine freedom was one of the flashpoints between falsafa (the Aristotelian camp whose champions were Avicenna and Averroes) and the mutakallimun (Islamic philosophical theologians, the prince among them al-Ghazali). The mutakallimun carried the day. In Christianity, one of the rare thinkers to deny divine alternatives was Peter Abelard.¹¹ He was condemned as heretical for it. As a result, Peter Lombard included an extensive anti-Abelardian section in his Sentences.¹² The Sentences became the standard theological textbook for centuries afterward. As a result, medieval Christian philosophical theologians cut their teeth showing in various ways why Abelard was wrong. Their work remains the basis for orthodox Catholic theology, and in this respect for Protestant theology as well. In the modern period, only Spinoza (if he counts as a theist) clearly rejected divine alternatives. Leibniz struggled to preserve them. He meant to do so, though opinions differ on his success.

Those able to do otherwise differ in how many options they have open to them on how many occasions. The more they have of each, the more leeway they have as free agents.

11 Voluntarism and God's Leeway

Descartes held that God is free to do whatever He has the power to do, and that His power could even make contradictions true (Kenny 1970, p. 150). For Descartes, God's reason places no constraints on what He might will. So His leeway extends as far as His omnipotent power. Call this extreme voluntarism. A more moderate voluntarist might e.g. hold just that God can sometimes choose against the balance of His antecedent motivations. The alternate approach is rationalism.

12 Rationalism and God's Leeway

The rationalist sees it this way. An omniscient being cannot miss the reason-giving force of any fact. God is perfectly rational. So if God has most reason to do a given act, He does it. There are objective truths about value. They are not an external constraint on Him. Rather, at least in Christianity, that His nature somehow embodies them has been by far the majority view since Augustine. Among those who deny that, the most popular view has traced them somehow to His will. As omniscient, God knows all these truths. So where there is an objectively best choice to make, He makes it. God automatically, by nature, as a sort of cosmic computer, tots up the relevant values and outputs an objectively best action if one is available. Abelard expressed this picture when he wrote, "to such a degree is God in all that he does mindful of the good, that he is ... induced to make individual things rather by the value of the good in them than by the choice of his own will" (Lovejoy 1936, p. 71). Leibniz had the same basic picture in mind when he wrote that God "is inclined toward all possible good, and ... this inclination is proportionate to the excellence of the good."¹³ On this picture, due to His perfectly rational nature, God has leeway only between equally good or incommensurably best results, or in infinite-hierarchy cases.

The Abelard-Leibniz “rationalist” picture of God’s freedom is alive and well. If emotions are ways to perceive value, Zagzebski holds one version of it. Swinburne thinks that “to believe that there is overriding reason to do it entails being inclined to do it, and doing it insofar as unimpeded by non-rational forces” (Swinburne 1994, p. 67). But for Swinburne, God is by nature subject to no non-rational influences (*ibid.*, 68, 128). Thus Swinburne concludes that

if there is a best action, he will do it; or, if there are alternative equal best actions, he will do one of them ... or a best kind of act ... God by his very nature will do it or one of them (as applicable). (Swinburne 1994, pp. 134, 135)

Rationalism vs. voluntarism is the great historic divide in discussion of divine leeway. But most recent discussion of divine freedom has centered on an argument that God’s goodness leaves Him almost no leeway. I discuss William Rowe’s version of this.

13 Rowe’s Argument

Rowe’s argument hinges on what he calls Principle B:

B. If an omniscient being creates a world when there is a better world that it could have created, then someone could be morally better than it is (Rowe 2004, p. 91).

(B) says in effect: moral perfection requires doing the best that is in one’s power. (B) speaks of creating worlds. But Rowe’s worlds are possible worlds in the modal-logical sense. So I state Rowe’s argument in terms of actualizing worlds.

It is standard perfect-being theism that

(1) God is omnipotent, omniscient, and morally perfect.

Add that

(2) God has actualized the actual world, and

(3) For every possible world God can actualize, there is a better one He can actualize.

Given (2) and (3), God has actualized a world, but could have done better. So per (B), someone could be morally better than God is. Per (1), God is morally perfect, and

(4) No-one can be morally better than someone morally perfect.

So God did not actualize this world (Rowe 2004, pp. 90–91). God’s goodness leaves Him unfree to actualize any world than which there is a better. So if there is no best possible world, no world is good enough for God to actualize. As there is none, God is not free to actualize any world. Further, if there were just one, God’s leeway would extend only to actualizing that. Further still, the argument can be extended to actualizing states of affairs “smaller” than entire worlds. It’s a short step from all this to God’s non-existence. Rowe takes that step, but that step won’t concern us here.

Here is a natural response to (B). If there is no best possible world, then no matter what, there could be a better. But that’s not God’s fault. One can’t be blamed for not doing one’s best

if there is no best to do. So we should not mark God down for this. This reply misses the point. Rowe is not blaming God. It's just that (allegedly) if He could have done better, He can't be perfect. A further reply then asks why He can't be perfect if He has no option but to choose some world than which there is a better. Perfection shouldn't require what can't be done. Rowe's answer is that perfection isn't compatible with being surpassable, and a being with a better action on its record would surpass one otherwise the same but with a lesser action.

14 Why (B)?

One might argue (B) in many ways, but here I discuss only Rowe's. Rowe argues (B) as follows: suppose that omniscient being 1 picks a world to actualize. 1 picks a world only if 1 judges it acceptable. What 1 judges good enough to actualize indicates how good 1 is. If 1 could have actualized a better, there could be another creator, 2, with higher standards than 1. 2 would rule the world 1 picks not good enough, and pick a better world. 2's standards would indicate how good 2 is. So 2's goodness would exceed 1's. So if an omniscient being actualizes a world though it could have actualized a better, someone could be morally better than it: which is (B).

I think this is too simple. Let us ask why 1 and 2 have the standards they do. Suppose that 2 has the higher standard just because 2 is fastidious. This is morally neutral. A standard had just because of a non-moral personality quirk implies nothing about 2's moral character. Suppose that 2 has the higher standard because 2 wants a world that makes people think well of it. This might be morally neutral too. It could also be due to vanity. That would be to 2's discredit. If it is, 1 might be better than 2 morally. So we need to add a clause about reasons for the standards. Another problem is that the worlds' betterness might not be moral. 2 might just pick a prettier world. Having better taste than 1 implies nothing about 2's moral character. So we get an argument for (B) here only if we specify that the better world is morally better. The real thought behind (B), then, is that

BETTER. actualizing a morally better world for appropriate reasons displays higher moral standards,

and there is always a morally better world to be had. Correspondingly, (B) is really

B*. If an omniscient being actualizes a morally lesser world than it could have, then there could be a morally better being.

But (B*) is too simple.

Suppose that God chooses between two worlds. W has 10 units of moral goodness, including the goodness of God's choosing W. W's creatures feel a total of 30 units of pain. W* has 8 of moral goodness (including the goodness of God's choice) and 10 of pain. Suppose that God chooses W* out of tender regard for creaturely suffering. Nothing about this choice suggests a morally lesser being. God's might well be the morally best choice. So what has a chance to be true is only that

B**. If an omniscient being actualizes a morally lesser world than it could have, and does so without proper reasons, then there could be a morally better being.

On its face, (B**) is going to be a lot less useful to Rowe than (B). For given (B**), the mere fact that there are morally better worlds does not paralyze God. Nothing about choosing the morally less entails that the choice was without proper reasons.

Further, BETTER and B** are ambiguous. There is more than one way to actualize a world: God strongly actualizes W iff He causes all states of affairs in W to obtain.

God weakly actualizes W iff He causes only some states of affairs in W, and for every one He does not cause to obtain, there is an S* in W that He does cause, and a true subjunctive conditional S* > S.

God probabilistically actualizes W iff He causes only some of W's states of affairs, and for every one He does not cause, there is an S* in W that He does cause, S* made it highly probable that S obtain, and S obtains.

God jointly actualizes W iff none of the first three is true, He causes some states of affairs in W to obtain, and creatures He does not fully control cause the rest.

If we disambiguate B**, we get

- B**1. If an omniscient being strongly actualizes a morally lesser world than it could have, and does so without proper reasons, then there could be a morally better being.
- B**2. If an omniscient being weakly actualizes a morally lesser world than it could have, and does so without proper reasons, then there could be a morally better being.
- B**3. If an omniscient being probabilistically actualizes a morally lesser world than it could have, and does so without proper reasons, then there could be a morally better being.
- B**4. If an omniscient being jointly actualizes a morally lesser world than it could have, and does so without proper reasons, then there could be a morally better being.

B**1 and B**2 seem plausible. B**3 does not. If God goes for probabilistic actualization, then even if He made it very likely that things turn out in a certain way, it was not up to Him whether they did so. That was left to others, or to plain randomness. He had (we can suppose) proper reasons for trying (we can also suppose) to get a different, better world. That God had good reason to take the risk of a lesser world does not entail that these would also be good reasons for that world to exist, if it did come about. If God only probabilistically actualizes a world, He should not get full credit for how things turn out. If He doesn't, a lesser outcome does not indicate a lesser preventing character – provided, of course, that He had adequate reason to only probabilistically actualize.

B**4 seems false. For if x and y jointly actualize a world, they each get full moral credit for only part of it. Suppose that x and y jointly actualize W, x and z jointly actualize W*, and W* is morally better than W. It doesn't follow that x displays better standards in W*. It could be that x does precisely the same thing in both, and the worlds' difference is wholly due to y and z. Further, if x's reasons for one world or the other turn out improper, that could be due to actions of y or z which were not up to x, against which x took all appropriate measures short of simple prevention (which we can suppose was not morally appropriate). So again, B**4 seems false.

Once we disambiguate “actualize,” there are four versions of Rowe's argument, involving four versions of 2:

- 2.1 God has strongly actualized the actual world.
- 2.2 God has weakly actualized the actual world.
- 2.3 God has probabilistically actualized the actual world.
- 2.4 God has jointly actualized the actual world.

If B**3 and B**4 are false, we can get a sound version of Rowe's argument only with 2.1 or 2.2. I now argue that 2.1 and 2.2 are false. If I'm right, Rowe's argument must be unsound.

15 On Strong Actualization (2.1)

On any version of (2), God actualizes the actual world – that is, things as they actually are. So He actualizes a world containing ourselves, and any freedom we have. If He strongly actualizes the actual world, God has inter alia an executive volition – the “let there be light” sort – that I do A. If He does, I cannot resist His will. So I must do A. Further, I do not initiate my actions: God does. God deterministically causes my actions. That is, necessarily, if God wills that I do A, I do A. No other future is possible, in that respect. I shall mean by theological compatibilism (TC) the thesis that God's deterministically causing a creature's action is compossible with the creature's doing it freely and/or with moral responsibility. Only theological compatibilists will think that 2.1 can leave my freedom and responsibility intact.

If TC is true, God can leave us free and responsible, yet causally determine that we do no moral evil. If He can, it is hard to argue that He has good reason to allow moral evil. On TC, defenses involving incompatibilist freedom are ruled out. It is hard, further, even to speak of His only allowing evil. On TC, He causes all of it. Thus if TC is true, very probably God does not exist. I think this is enough to show that theists cannot accept TC. This is so a fortiori if the theist holds any version of Western monotheism that allows a hell that is everlastingly populated. I now argue against TC another way.

16 Theological Compatibilism?

Let us call compatibilism the thesis that some real feature of the world that makes determinism true does not undermine or remove freedom and/or moral responsibility. Even if compatibilism is true for some determinism-inducing features, TC might be necessarily false. That some form of determinism does not (say) undermine moral responsibility does not entail that divine determinism does not. On ordinary physical compatibilism, natural causes bring about the internal states that make me (supposedly) responsible for my actions. No responsible agent causes this. So there is no-one to whom to transfer my moral responsibility, and so (physical compatibilists argue) it remains mine. On the other hand, if I cause you to commit murder by manipulating your internal states, there is a strong intuition that responsibility transfers to me. I am to blame for the murder, not you (Mele 2019). Divine determinism, I now suggest, inherits this intuition.

Divine determinism might well work this way: God sets up the world's causal order. This determines our initial physical condition and all later influences upon us. These make us evolve as we do. God sets up our internal states and environment so as to bring about the conduct He intends for us. In effect, then, He uses nature to manipulate us. His doing so makes us (in effect) His tools to bring about results He desires in the world by way of our intentional agency.

Tools are not morally responsible for the way they are used (Todd 2018). Suppose that my saw was a conscious intentional agent, and my way of getting it to cut wood was to telepathically influence its mind so that it moved as I wish. In this case I move the saw by mind rather than muscle, but still it moves only as I design. It is strongly unintuitive that the saw would in

this case be free or responsible. This is so just because the ultimate intentional explanation for its activity does not involve its own intentions at all. Though I use its mind, the real explanation for what it does lies in my mind. If one explained the saw's activity by a story about the saw's mind, then later learned of my role, one would think that first story had been true but not the real explanation. But so then in God's case, on divine determinism. His is the operative intent and moral responsibility. They usurp ours. Thus even if some form of determinism does not remove our moral responsibility, I submit that TC is false.

I am not nature's tool, because there is nothing nature is trying to get done by molding me. But if another will is involved, it is imposed on the manipulated agent. The manipulator compels the manipulated agent to have his/her mental states, as may not be true when no-one is there to do the compelling. This sort of compulsion undercuts moral responsibility. If we ask why the act was done, we have to look to the manipulator's intention, not the manipulated agent's. The manipulated agent's own intentions cease to be morally significant, though they continue to be causally so. We can only praise or blame the manipulator. Thus while divine determinism is certainly a possible view, TC is not.

17 Moral of the Story

If God strongly actualizes the world and TC is not true, what is true instead is divine hard determinism: God causes everything and there is not even a compatibilist version of creaturely freedom or moral responsibility. Divine hard determinism cannot appeal even to values achieved by compatibilist versions of creaturely freedom to justify God's causing the evil He has. So the probability of no God on a hard-determinist version of 2.1 is even higher, I think. Turning things the other way round, if there is a God, the probability of 2.1 is extremely low. If there is no God, on the other hand, 2.1 is false. Thus 2.1 is almost certainly false.

18 Weak Actualization: Molinism (2.2)

2.2 is the standard Molinist story about God. Before all creation, there are counterfactuals of freedom (CFs), true conditionals of the form $S > S^*$ in which S^* is some creature freely performing an action. God knows these. He uses them to set up the history He wants, including all free choices He foreordains. So I could just lean on standard criticisms of Molinism to reject it. I think these really suffice. But I'll add a point: Molinism is a form of TC, and so falls under my previous argument.

A standard account of determinism has it that a world is determinist at t just if, given the past up to t , there is only one way that world's future can go: its future is closed, not open. This is true on Molinism. Given the CFs and God's eternal decision on how to set the world up initially, there is only one way the future can unroll, though there are many ways it could have unrolled (had God eternally made a different use of the CFs). So Molinism is determinist. Thus if Molinism includes any sort of freedom at all, it is a form of TC, just by definition. Its distinctive that it is a non-causal TC. God does not have an executive volition that I do A, on Molinism. Instead, God executively wills that S obtain, knowing that $S > I$ do A and intending by so willing that I do A. There is no causal compulsion here. But still, the ultimate responsibility for my doing A goes to God here, even if He doesn't cause my choice. So 2.2 is out, and so Rowe's argument fails.

19 On 2.3

I add that 2.3 seems unlikely. 2.3 would work this way: God determines some contingent facts, including contingent conditions which make it likely that the rest work out as He wants, and then lets other agents bring about the rest of the desired world (e.g. by their free choices). While it is not impossible that this get Him the world He wants, the odds against it are daunting. Suppose that God so stacks the probabilities that every contingent state of affairs He wants but leaves up to others has a .9999999999 probability of coming about. Then if He leaves enough desired contingent states of affairs to others, the probability of getting them all sinks as close as you like to zero – and it takes getting them all just right to actualize a possible world He wants. So God can follow this method and be fairly sure to get what He wants only if He leaves very little in the hands of (say) very few created free agents whose freedom is greatly restricted. The more free choices are left open in a world, and the less He biases the system toward the results He wants, the less likely He is to get them by this means. I note too that if the probability of the world He wants sinks too low, He should not really get credit for actualizing it even probabilistically. For no-one should be credited with merely lucky consequences of their actions.

It seems unlikely that God has tried to get just the world He wants in this way. For it seems to us that a great deal has been left to created free agency in our world, and holding that it has is part of free will defenses, without which the problem of moral evil is very hard to handle. So I suggest that theists can safely disregard 2.3. If God has created free agents and left things in their charge, He has not attempted by doing so to actualize precisely one possible world. He would have known that the odds against getting just that result were daunting. As perfectly rational, He would not have tried.

I think, then, that 2.4 is true. Here is its implication for Rowe. Suppose, for instance, that God sets in motion indeterministic processes He does not control, and these affect how much good He winds up doing, and which possible world winds up actual. Assume He does His best to direct the processes, short of controlling them, which we're assuming He has good-enough reason not to do. Then things God makes might well bring about less good than they might have. If they do, God winds up doing less good overall (directly, by creating, and indirectly, by setting these processes in motion) than He might have. He gets a lesser world than He would have preferred. But this is compatible with His goodness' being perfect if He had good-enough reason to proceed in this way and did not determine the outcome.

Notes

- 1 For one such view, see Lewis (1973), pp. 73–77, (1999), pp. 8–55, 224–247.
- 2 An exception is Roberts (2008).
- 3 See Leftow (2009).
- 4 I do not know whether Jewish or Islamic thinkers ever raised the issue.
- 5 There is a folk tale that Peter Damian disagreed. I think this is a misreading, but I cannot pursue this here.
- 6 See e.g. Aquinas, *Summa Theologiae* Ia 25, 4.
- 7 If determinism is true, then it is part of the one fixed history that God overrules and unmakes it.
- 8 They are not preferences themselves. No attitude just to P is a (rational) preference that P. A rational preference involves at least two claims: I prefer its being the case that P to its being the case that Q. If it is possible to prefer that P to that P – I take no stand on this – such a preference is irrational.
- 9 Swinburne denies that He does: see Swinburne (1994, pp. 68, 128).

- 10 Thomas Senor notes that if such things as this are true, then even when God can't refrain, as when He keeps a promise, only this inability will be any reason to call His action unfree (Senor 2008, p. 184).
- 11 For texts and discussion, see John Marenbon, *Abelard in Four Dimensions* (Notre Dame, In: University of Notre Dame Press, 2013).
- 12 Bk. 1, d. 43.
- 13 Leibniz, Theodicy, #80.

Bibliography

- Boethius. (1973). *The Consolation of Philosophy* (ed. H. F. Steward, E. K. Rand and S. J. Tester and tr. Boethius). Cambridge, Mass.: Harvard University Press.
- Bourne, C. (2006). *A Future for Presentism*. Oxford: Oxford University Press.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control*. Cambridge: Cambridge University Press.
- Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5–20.
- Hudson, H. (2014). *The Fall and Hypertime*. Oxford: Oxford University Press.
- Kane, R. (1996). *The Significance of Free Will*. Oxford: Oxford University Press.
- Kenny, A. (ed. and tr.) (1970). *Descartes: Philosophical Letters*. Oxford: Oxford University Press.
- Leftow, B. (2009). Omnipotence. In: *The Oxford Handbook of Philosophical Theology* (ed. T. Flint and M. Rea), 167–198. Oxford: Oxford University Press.
- Leftow, B. (2012). *God and Necessity*. Oxford: Oxford University Press.
- Lewis, D. (1973). *Counterfactuals*. Cambridge, Mass.: Harvard University Press.
- Lewis, D. (1999). *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.
- Lovejoy, A.O. (1936). *The Great Chain of Being*. Cambridge, Mass.: Harvard University Press.
- Mele, A. (2019). *Manipulated Agents*. Oxford: Oxford University Press.
- Merricks, T. (2007). *Truth and Ontology*. New York: Oxford University Press.
- Pruss, A. (2013). Omnipotence. *Res Philosophica* 90: 1–21.
- Roberts, J. (2008). *The Law-Governed Universe*. New York: Oxford University Press.
- Rowe, W. (2004). *Can God Be Free?* Oxford: Oxford University Press.
- Senor, T. (2008). Defending divine freedom. *Oxford Studies in Philosophy of Religion* 1 (2008).
- Swinburne, R. (1994). *The Christian God*. Oxford: Oxford University Press.
- Todd, P. (2018). Does God have moral standing to blame? *Faith and Philosophy* 35: 33–55.
- Wolf, S. (1990). *Freedom Within Reason*. Oxford: Oxford University Press.
- Zagzebski, L. (2004). *Divine Motivation Theory*. Cambridge: Cambridge University Press.

Denialism

SAUL SMILANSKY

Denialism (on free will and moral responsibility; henceforth the disclaimer will be dropped) in its minimal version combines the first two among the following philosophical positions. A broader version of denialism combines all of the following five philosophical positions.

- (1) NON-LIBERTARIANISM: There is no libertarian free will (LFW) and hence no LFW-based moral responsibility
- (2) INCOMPATIBILISM: Compatibilism with respect to the possibility of free will and moral responsibility in a world without LFW (such as a deterministic world) is mistaken
The combined implication of (1) and (2) is that there is no free will and moral responsibility; of either libertarian or compatibilist kinds. This, again, suffices for a minimal form of denialism. A broad version adds the following three positions:
- (3) THE AWFULNESS OF THE STATUS QUO: The belief in free will and moral responsibility and the connected reactions and practices are, all considered, extremely harmful
- (4) THE VIABILITY OF ALTERNATIVES: There are good and realistic alternatives to the beliefs in free will and moral responsibility and the connected reactions and practices
- (5) WE SHOULD OPT FOR RADICAL CHANGE: We should, all considered, aim to bring about a radical change; and to live without the beliefs, attitudes and practices which assume libertarian or compatibilist senses of free will and moral responsibility

I have chosen the term “denialism.” Some use the terms “hard determinism,” “hard incompatibilism,” or “free will skepticism” for roughly this sort of view, but each such possibility has different unwanted connotations. There is a need for some tolerance as to the status of libertarian free will: all denialists will of course deny that it exists, but we need not decide here whether it makes sense or is incoherent, for example, and as we shall see, there are other differences among denialists in their attitudes.

The free will problem is (a) indisputably one of the great problems of philosophy and has been known for at least two thousand years. It is (b) a complex problem which includes at least five distinct major questions (Smilansky 2017a). (c) The last two generations have been

a golden period for the free will debate and arguably more progress has been made during those years than in the previous two millennia; and (d) this is probably even more so for denialism, as denialists have become much more numerous, active and important philosophically in recent years. The combination of these four factors is that there is an enormous literature that comes within our sphere of concern. I will be selective, depending on what I hold to be the most important issues and views, in a way that necessarily reflects my own views and limitations. I apologize for leaving out so much.

Given these broad complexities, it is fortunate that we can make some assumptions. First, we shall assume Position 1 above, i.e. that libertarian free will does not exist, either because determinism is true, or because the sort of indeterminism that is true cannot generate LFW, or because, independently of determinism or indeterminism, the very notion of LFW is incoherent. I refer to “robust” LFW to exclude libertarian views such as those of Kane (1996), which famously lacks the control element, and hence cannot give us much beyond what compatibilism can provide.

Second, we shall be assuming that free will and moral responsibility go together. This is in broad agreement within the free will debate (the major exception is Waller, who believes in free will but not in moral responsibility; some attempts have also been made to ground MR without FW, but these need not concern us here). Most of the interest in free will derives from the implications for moral responsibility and related notions, although not all (for example, concern about meaning in life). And the notion of free will that is nearly always the focus of debate in the free will problem is that which is required for moral responsibility.

1 The Nature of the Disagreement on the Compatibility Question

It seems that concerning the compatibility question (as in Position 2 above), there is a further widespread agreement between compatibilists and denialists, which sometimes needs uncovering, and that has not been as widely recognized as the two points above. This agreement is highly important, although it does not, of course, prevent denialists and compatibilists from fundamentally disagreeing on the compatibility question itself. The broad and considerable agreement concerns the *descriptive aspects* of the topic; the facts, in a sense. The disagreement, regarding the compatibility question, is essentially about the *interpretation* of these largely agreed-upon facts. This can be readily seen from the traditional terms “soft determinism” and “hard determinism” (which originated with William James 1896/1956). These can adequately stand in for compatibilism and denialism, although today both sides need not be committed to the universality of determinism, for indeterminism as such is of no particular interest to (and should not be thought to decisively benefit) either compatibilism or denialism. It is a common error for beginning students of the free will problem to think that hard determinists believe in a different, “harder” sort of determinism than compatibilists, but this is as a rule false. Both are equally determinists yet disagree on the interpretation, namely, on whether this rules out free will and moral responsibility and the concomitant notions.

Consider an example of a person who is trying to choose what to eat for lunch, the options being schnitzel or spaghetti. Both compatibilists and denialists will agree that there is a situation of choice here; as a rule, denialists do not deny choice. Both would further agree that a typical adult person could choose either schnitzel or spaghetti, in the sense that he would do so, if he preferred one or the other. The disagreement is that traditional compatibilists, at least since the time of Hume, have claimed that this is the sense of free will that matters, often adding that there is nothing more here that we could wish for. Denialists, by contrast, have claimed

that this is not at all the sense of freedom that matters, but what matters is something like whether the person (assuming he chose schnitzel) could have in fact chosen otherwise, i.e. opted for spaghetti, as he and the world were, i.e. with internal and external conditions held constant. But note that here again the compatibilists have not wanted to disagree that, indeed, if the person was exactly as he was and situated in exactly the same causal nexus, the schnitzel-choice would emerge. The compatibilist may or may not be a determinist, but either way the causal determination of the schnitzel-choice does not bother her; she only insists that, as long as the person wants schnitzel and chose as he wanted, he is completely free (and presumably responsible for that choice).

Of course debate on the compatibility question has become extremely sophisticated, and has gone much beyond the two contrasting interpretations of the “ability to do otherwise.” Refined compatibilist interpretations of the requirements for freedom and responsibility have specified the conditions under which agents are indeed free and responsible beyond mere consistency between what they happen to desire and what they choose. This is a good idea, given that sometimes agents are clearly not free precisely because of what they desire, although they do go on to choose what they desire. If the schnitzel-chooser of the previous example, on being informed that today there are excellent reasons to pass it over, would still insist on having it, he might be naturally suspected of not being free (assuming he does not wish to get sick). Likewise, in a paradigmatic case, a woman who desires to wash her hands every few minutes will not be deemed free even if she does indeed wish to do so. Compatibilists have suggested various ideas, such as hierarchical models (e.g. Frankfurt 1988), reasons-responsive ones (Fischer and Ravizza 1998), or rational-abilities-based views (e.g. Nelkin 2011), as ways of explicating compatibilist freedom. Note that here again the denialists will not want to deny that the compatibilist distinctions capture *a* sense of freedom – denialists do not deny that we are better off without kleptomania, alcoholism, or other compulsions. It is only that while the compatibilists think that matters more or less end there, the denialists insist that something important is missing. There are further, more demanding conditions for free will and moral responsibility, which (given the absence of LFW, or indeed the impossibility of any robust sense of free will) cannot be met.

Another way in which we can see this is by looking at the notion of control, which is typically a stand-in for free will throughout the free will debate. Compatibilists often speak about self-control and, by this, they mean things such as the ability, through reflection, to increase one’s awareness of the choices available; to better evaluate the reasons one has for choosing this or that choice; to carry through with one’s preferences, and the like (as in the accounts above by Frankfurt or by Fischer and Ravizza). Denialists, by contrast, think of the morally required self-control in a much more demanding way, as requiring self-creation, or ultimate control. This would not rule out versions that think that free will does not make sense; the sense of “morally required” is just a way of saying that compatibilism is morally insufficient. According to Galen Strawson’s Basic Argument, for example, “(1) Nothing can be *causa sui* – nothing can be the cause of itself. (2) In order to be truly morally responsible for one’s actions one would have to be *causa sui*, at least in certain crucial mental respects. (3) Therefore nothing can be truly morally responsible” (Strawson 1994, p. 6). Waller likewise holds that “Just deserts and moral responsibility require a godlike power – the existential power of choosing ourselves, the godlike power of making ourselves from scratch, the divine capacity to be an uncaused cause – that we do not have. Moral responsibility is an atavistic relic of a belief system we (as naturalists) have rejected, for good reason” (Waller 2011, p. 40).

I cannot review here the manifold arguments that denialists have presented against compatibilists on the compatibility question and the opposing moves of their opponents.

But the common feature is that denialists put the bar much *higher* than compatibilists, and think that free will and moral responsibility (if it were at all possible) would require that agents jump over that bar, while this is conceptually or practically impossible. Denialists put the bar where libertarians do but, unlike libertarians, are pessimistic that agents can jump as high. Compatibilists, by contrast, put the bar for free will and moral responsibility much *lower* and hence, for them, most people most of the time (i.e., normal adults under typical conditions) can clear it, and be considered sufficiently free and morally responsible.

2 Absolute Denialism and Partial Denialism

This interpretation of the compatibility question helps us to see an important distinction about denialism, that between *Absolute Denialism* (AD) and *Partial Denialism* (PD). AD denies that it is in any way or to any extent adequate to believe in free will and moral responsibility; it is a “high-bar” belief. PD holds that compatibilism, the “low-bar” position on the conditions for free will and moral responsibility, is too simplistic, and indeed perhaps complacent; since even where compatibilist conditions are met, there are often further reasons to see difficulties concerning free will and moral responsibility. Yet PD holds that compatibilist freedom and the resulting potential for moral responsibility are valid and important, that they do capture some of free will and moral responsibility, but they just come up short. AD says that compatibilist conceptions of free will and moral responsibility completely miss the point. PD, by contrast, thinks that the compatibilist conception of freedom doesn’t completely miss the point and can even capture some true insights about moral responsibility, but just not all of them. So on Position 2 PD is only in partial agreement, as it shares the insights of denialism while giving some weight to compatibilism too.

Throughout its history the compatibility question on the free will problem has been conducted under the Assumption of Monism whereby we can be either compatibilists or incompatibilists (and hence, given the absence of libertarian free will, either compatibilists or denialists). Yet this standard assumption is false (Smilansky 1993, 2000, 2003); we can combine some measure of both. This means that one can be a partial denialist; holding that compatibilism is not totally mistaken, yet that the absence of LFW greatly matters, and requires considerable philosophical revision and practical reform (or, indeed, revolution). The importance of this “dualistic” or pluralistic possibility has not been sufficiently internalized in the debate, and hence those inclined towards denialism often just assume that they cannot hold this view in moderate form, and must be absolute denialists. An early anticipation of compatibility dualism or pluralism is Jonathan Glover’s unduly neglected “Self-creation” (1983). Ted Honderich (1988) proposes an “emotive” account which affirms compatibilist-like attitudes within a framework skeptical of traditional free will views. Martha Klein (1990) considers the limitations of compatibilism yet partly affirms it. Paul Russell (e.g. 2013) offers a “pessimistic” compatibilism which is aware of its inability to fulfill traditional optimistic compatibilist claims. Another way in which the standard Assumption of Monism has been rejected is that of contextualism (e.g. Hawthorne 2001; Feldman 2004). Contextualists hold that one ought to be a compatibilist or denialist in response to the context: in some contexts the appropriate attitude towards the compatibility question is the former, and in other contexts it is the latter. The most extreme pluralistic position on the compatibility question has been proposed by Saul Smilansky (Smilansky

2000: Part I, 2011a), who shares the contextualist view that sometimes we ought to be compatibilists and sometimes denialists, but adds that to some extent we ought to be both at the same time, so that at least on some occasions a view trying to combine the limited insights of both contrasting positions together is the most convincing. Neither compatibilism nor denialism is, according to this view, the whole truth, but both capture part of the true, inherently complex, picture.

3 The Status Quo

The belief that there is free will and moral responsibility is very common, as are the attitudes and practices connected to it. We can begin with what we can call the Core Conception of beliefs concerning free will, moral responsibility, and the concomitant beliefs, reactions and practices (Smilansky 2000: ch.2, 2022). This is quite general and often quite vague, as well as an idealized version of prevailing views, which are often not so pure. According to this Core Conception, free will or control is the basis for moral responsibility, which is in turn a condition for deserving many sorts of evaluations, reactions, and treatments. Self-attitudes, interpersonal relations, social interactions, institutions, and practices should all, in manifold ways and within constraints and limitations but nevertheless very broadly, *respond to* and *track agency and responsibility-relevant free control*. This is moral responsibility of the desert-involving kind, not only a forward-looking one. Hence, for example, pro tanto we ought to create social orders in which personal choice and responsibility can be widely exercised and rewarded, accept moral responsibility and reasonably hold ourselves accountable for our negative free actions, be grateful to those who freely help us, appreciate people (including ourselves) who make significant sacrifices or take risks for good moral reasons, and take great care and even risks in order not to blame or punish the innocent who lacked control over their actions.¹

Both libertarians and compatibilists assume the Core Conception, and think that it can and ought to guide our lives. As we have already noted, libertarians, like denialists, are “high-bar” but unlike the denialists they are optimistic that there can be conceptual and metaphysical grounding for it, namely robust libertarian free will.² Compatibilists are likewise optimistic but this is because they are “low-bar,” and think that compatibilist forms of free will and moral responsibility suffice. Denialists, at least in the absolute form of denialism, are here pessimistic, because they deny both libertarianism and compatibilism. There can be a denialism that would stop here. It would agree with the libertarians and compatibilists about the importance of the Core Conception, but deny its application. We can call this “*Depressive denialism*.” This form of denialism actually seems philosophically attractive, although in itself could hardly form the basis for free will–related personal and social life.

Most contemporary denialism has not been of this form, but has been fairly revolutionary and very optimistic. It has rejected the Core Conception. This has involved going beyond the minimal denialism of Positions 1–2 (the rejection of LFW and compatibilism), and holding Positions 3–5, namely, the awfulness of the status quo, the viability of good alternatives, and the conclusion that we should opt for radical change. This embodies a “good-riddance” attitude, which involves *offensive* and *defensive* aspects of denialism. The offensive aspects aim to show why we would be better off without belief in free will and moral responsibility, and is based upon arguing for the faults of the *status quo* (Position 3). The beliefs in free will and moral responsibility, and the associated attitudes and practices are so harmful, it is held, that every effort should be made to give up these beliefs.

4 The Awfulness of the Status Quo according to Denialism (Position 3)

Not surprisingly, the major focus of the denialist offensive has come out concerning punishment, and here as well we shall focus mostly on it. Punishment is a good topic for moral responsibility denialists because it is an area where unnecessary evils are easy to show. Particularly in the US, which is the denialists' almost exclusive focus, there is clearly much public retributivism gone wild, over-criminalization, over-punishment and, simply, cruelty. Contemporary practices of retributive punishment can easily seem antiquated, unnecessarily vindictive, brutal, and primitive.³

Most contemporary denialist discussions have focused on punishment and on related beliefs and reactions such as blame. Pereboom emphasizes the moral and psychological vile-ness of retributivism, while “[T]he good by means of which retributivism justifies punishment is that an agent be the target of harmful treatment just because of his having knowingly done wrong. This position would be undermined given free will skepticism, because if agents do not deserve blame just because they have knowingly done wrong, neither do they deserve punishment just because they have knowingly done wrong” (Pereboom 2014, p. 157). Waller similarly says that “We must not become comfortable with punishment. That is the great wrong of moral responsibility. When we can appeal to moral responsibility and just deserts, punishment becomes not only morally acceptable, but morally required; not a disturbing problem, but a positive good; not a troubling misfortune, but a celebration of personhood; not a deeply unfortunate wrong, but righteous retribution” (Waller 2018, p. 20). Moreover, “[I]t is the belief in moral responsibility itself that is causing many of the worse problems, and real reform will be facilitated by dropping or at least minimizing the commitment to just deserts and moral responsibility” (Waller 2018, pp. 16–17).

5 The Viability of Alternatives (Position 4)

5.1 *Punishment*

The topic of punishment is also crucial for moral responsibility denialists in their defensive mode. The paradigm Core Conception of our thinking about the free will problem sees individual control as the basis for moral responsibility which is, in turn, the basis for the treatment and reaction we deserve for our good or bad actions. That is why punishment of the innocent, of the very young, or collective punishment, for example, are a moral anathema; they are paradigms of injustice. So what do the denialists have to offer instead? Can they maintain the old Blackstonian moral idea that it is better that ten guilty persons escape than that one innocent party suffer? But how, when for denialists *everyone* is morally innocent, whether she has done anything or not? More broadly, can denialists resist the constant temptations for the efficient management of people? Can they respect justice and other values that have been dependent upon the idea of moral responsibility and, for example, avoid joining the utilitarians on punishment?

Some alternatives to current practices of punishment will need to be introduced instead, and this relates to the defensive aspect of denialism, where the denialists need to defend against claims that giving up belief in free will and moral responsibility would be extremely dangerous. Both aspects are necessary in order to make the “good riddance” claim of denialism. For it is not enough to argue that belief in free will and moral responsibility currently does much harm, unless a plausible case can be made concerning the replacements.

The most detailed attempt to provide a denialist alternative to the Core Conception on punishment is Waller's recent book (Waller 2018). His main proposal is a no-blame system model that originates in the workplace. "Rather than *blaming* or shaming an individual worker as the source of the problem, the system model treats errors and mistakes – whether large or small – as learning opportunities that can improve the workplace and the production system. A worker who reports an error ... is thanked rather than blamed" (pp. 163–164). The emphasis on individual fault and blaming for it is transformed into a constructive, open search for the causes of failure beyond individuals, to which the individuals are encouraged to contribute.

However, Waller's non-blame systems approach does not address the predicament of punishment. There are good reasons to recognize the advantages of this approach to management in areas as diverse as airline controllers, vehicle construction, and perhaps even some parts of medical practice. Broader attempts at their social application would be welcome. But when it functions well, the non-blame culture, by and large, works on the assumption that the persons involved are people of good will, eager to work together for the common cause. That seems to assume away any realist assessment of most serious criminals. Engaging Mafia soldiers in such talk about "failure" seems an obvious non-starter. It is not an accident that the no-blame approach typically deals with at worst negligence, rather than outright malevolence and intentional destruction. Moreover, any psychological description of the good members of such communities seems to me to take for granted that they do hold themselves responsible, and would blame themselves were they not to meet their own or the group's standards. The no-blame culture seems a welcome, if limited, addition to the way in which moral responsibility-accepting persons will engage with themselves and others.

A more sustainable denialist approach to punishment uses the analogy of the quarantine model, as offered by Pereboom (2014, among others), and the combined quarantine and medical model offered by Caruso (2016). As Pereboom says, "Suppose someone poses a danger by threatening to commit murder. Even if he is not in general a morally responsible agent in the basic desert sense, the state would nevertheless seem to have as much right to isolate him as it does to quarantine a carrier of a deadly communicable disease who is not responsible in this sense for being a carrier" (Pereboom 2014, p. 169).

There have been diverse forms of criticism against these radical revisionist ideas of denialists. One major difficulty concerns the unjustness of punishment for denialism. I have argued that from the denialist perspective no one can be *justly* punished, and hence cannot be effectively threatened with punishment. At most, denialists can opt for "funishment," incarceration joined with very generous living conditions (in order to compensate for the undeserved injustice of incarceration and the denial of freedom involved). But this direction is then self-defeating for denialism, encouraging rather than discouraging more people to risk engagement with crime, and therefore causing more incarceration and in general making things worse *from the denialist perspective itself* (Smilansky 2011b; for criticism of my "funishment" argument see Levy 2012; Waller 2015, pp. 197–200; Pereboom 2017a). Not to recognize this would mean betrayal of the negative but true insights of a morally deep denialist position, concerning the moral baseline of universal innocence and the concomitant injustice of punishment.⁴

Such a direction also involves denialism with the traditional grave dangers of such forward-looking consequentialism (see Smilansky 2017b, 2019). Denialists have emphasized the drawbacks of traditional views which follow the free will paradigm of the Core Conception, but these views are often also safeguards of equity, decency, and human rights. One example concerns punishment of the innocent. Criticism of utilitarianism here has traditionally focused on

extreme examples such as scapegoating. But as I showed long ago, the real danger concerns a systematic, moderate lowering of standards for prosecution and conviction (Smilansky 1990, 2000, pp. 28–30, with relevance to free will; see also the development of this argument in a critique of the dominant denialist quarantine model by Lemos 2016). The move of focusing on the ineffectiveness of the threat of punishment on the innocent themselves will not do (see Smilansky 2000, pp. 29–30). Consequentialists at least since Bentham counter that consequentialism has its own resources to limit dangers. As Levy puts it, “A consequentialist who is a moral responsibility skeptic will naturally hold that no one should be treated any worse than is needed to bring about the best consequences, with all agents’ welfare – including the welfare of criminals – taken into account” (Levy 2012). But this will comfort only consequentialists and, as we shall see, many denialists have sought to defend deontological constraints.

A parallel set of risks and dangers has long been familiar concerning the medical analogy for substitutes to punishment. The idea that wrongdoers are merely ill, and their wrongdoing needs to be treated (rather than their being blamed and punished), may seem progressive and humane. But as has been often argued (e.g. Hart 1970; Murphy 1973; Morris 1976), this is too quick. Free will–based models of punishment have the resources to limit punishment to actual wrongdoers (bracketing epistemic issues). They are also typically constrained in their punishment by the proportionate moral “price” that the given crime merits, so that one will be punished only for, and to the degree of, the given crime that one committed and hence deserves to be punished. And there are severe limitations on the permission of breaching the integrity and autonomy of the punished.

Medical models are not so constrained. Since the main concern is the forward-looking “healthy” social functioning of the potentially incarcerated, they could be deprived of their freedom even if they have not done anything. Unlike the free will–based model, with the medical model, once one is in, there is no knowing in advance as to when one will be released. Finally, since the concern is with changing the ways of the incarcerated, constraints over behavioral and medical treatments are much weaker. All this pertains to the legitimate practices according to the medical model; but this model also raises considerable opportunities for “Cuckoo-nest-like” abuse. One need not be blind to the wrongs and horrors of the current justice system to see that the medical model poses special wrongs and risks. Corrado (2017) and Waller (2018) are two denialists that have been particularly aware of the dangers of the medical model and its temptations.

Denialists have attempted to defend the idea of deontological constraints in two major ways. For Pereboom, “The concern about using people merely as a means has force in this context, and this together with the weight of the general right to liberty should restrict general detention to especially dangerous cases” (Pereboom 2014, p. 169). Punishing the innocent would be excluded because, not being offenders, they pose no threat to us. Similarly excluded would be punishing offenders more than is strictly necessary for defending ourselves from them. Vilhauer (a denialist due to epistemic skepticism about free will and moral responsibility) has a different approach here: he claims that the idea of contractual consent, in the familiar deontological-Kantian way, can justify punishment, while at the same time provide a constraint on using some for the means of punishing others, in a way that should allay our worries about utilitarian-like abuse. The way to do so, Vilhauer suggests, is to follow Rawls: “we can use original position deliberation to develop a justification of punishment which has a ready reply to the mere means objection, but which can nonetheless be endorsed by free will skeptics” (Vilhauer 2013, p. 153).

In response to both, I presented seven reasons for skepticism about optimistic free will denialism, concerning the stability of deontological constraints in a world without free will

and moral responsibility (Smilansky 2019). Here I can just note them in brief. The effort to establish the deontological constraints has been a huge historical struggle, one of the greatest moral achievements of humankind; and it has been achieved through the emphasis on agency and individual fault. Moreover, we have no real idea how a society which does not hold its mature members responsible will look. Once we examine the nature of the relevant deontological constraints and the motivation behind them, they seem to be inherently attached to notions of free agency and (non-universal) innocence. The ideas of denialists become under-motivated without the free will–related desert background. Looking at distributive justice, where agency is not taken very seriously in many important contexts (e.g. inheritance dominates), gives us further reason to worry. There is a big question mark as to whether the arguments of the denialists are philosophically compelling, e.g., about Rawlsian contractualism as interpreted by Vilhauer. But once we see that denialism has a further mountain to climb, the pragmatic one, and the pragmatic weakness of what denialism has to offer, the optimism about the consequences of adopting denialism becomes even less persuasive. Finally, these points are strengthened once we realize that a major challenge concerns a statistics-based lowering of standards, which we already noted, rather than scapegoating or similar worries. This makes it all the more difficult for denialism to safeguard the deontological constraints, in a way that Core Conception beliefs about the primacy of agency, guilt, innocence, and responsibility-based desert typically can.

The call for efficiency often suffices to sidetrack justice and decency. If our focus shifts away from concern over freedom and responsibility, and we are no longer concerned with giving people what they deserve (and hence making a big effort, and taking risks, not to punish the undeserving), then the fundamental deontological intuitions and constraints are likely to lose most of their attractiveness and influence. There is a danger that people will treat one another in large measure as though they are mere carriers of features or “symptoms” that are to be dealt with, rather than treating one another as agents capable of reasoned choice and responsibility, who should be responded to and treated according to their choices and actions.

There is a close connection here between the Core Conception concerning free will and moral responsibility, and respect for persons (Smilansky 2000: section 2.1, 2005; see also Bradley 1927: essay 1; Hart 1970; Morris 1976; Berlin 1980). To respect persons is to take seriously their agency, their choices and actions in themselves, not just for consequentialist reasons, to see them as autonomous and responsible agents who are appreciated and treated in accordance with their intentions, efforts and actions.

5.2 *Punishment, Justice and Justification*

In the light of discussions such as those which we have surveyed, there is considerable disagreement among denialists as to whether punishment can be just. As we saw, Pereboom and to some extent Caruso believe that it can be, although of course efforts also need to be made to change the environments that lead to crime and therefore to reduce the need for punishment. The quarantine model, which is largely based upon the idea of legitimate social self-defense, and the belief that deontological constraints can be maintained alongside a largely forward-looking practice, come together to form a view that denialist practices can be largely just. With Vilhauer, as we saw, the justification is contractual, and he holds that it can override independent Core Conception–like intuitions about justice. And Parfit (2011) believes that the idea of fair warning can serve to make justice fair and just, even under denialist assumptions (cf. Smilansky 2016).

Other denialists do not hold that punishment can be just. Corrado says that “If the hard incompatibilists are right and no one ever acts freely, then punishment is never justified: punishment would always be punishment of the innocent” (Corrado 2018, p. 64). Waller believes that punishment as such cannot be just: “We may label the punishment with a new name: preventive detention, therapeutic intervention. Whatever we are doing, we should not hide from ourselves whatever it is that we are doing. Some criminals *must* be detained for the protection of society, and they do *not* justly deserve such treatment. Painful as it is to be involved in such injustice, we should acknowledge that an injustice is being done” (Waller 2018, p. 91).⁵

In my view denialists need to admit that punishment is unjust at least in a central sense, which has been the focus of the free will problem (it might also be thought possible that it can be just in some other senses). And, as a part-denialist, i.e. a compatibility dualist, I do admit this. There is no substitute concerning justice for the paradigm of the Core Conception, when we are exploring the harsh measures of even the more progressive forms of the incarceration of dangerous criminals, and free will and moral responsibility-based desert is a condition for just punishment. If denialists deny free will and moral responsibility, then they must acknowledge that the punished are morally innocent, and to inflict on them anything like punishment is, in a fundamental way, unjust. In other words, they are being punished for constitutive and other forms of luck (Nagel 1986, 2012; Smilansky 2000, pp. 45–47; 2003; Levy 2011), for matters ultimately beyond their control. This question of injustice needs to be differentiated from the question of justification, and here there is room for a large variety of views. Diverse moral concerns mingle with pragmatic ones, and there are various options for the moral justification of practices that resemble punishment. There is also ample room for further work here.

5.3 *Blame Substitutes, the Reactive Attitudes, Appreciation, Respect, and Meaning in Life*

Beyond the issue of punishment, hard determinists have combined two approaches. The first can be described as “obdurateism,” claiming that more or less the same reactive attitudes and common emotions can be sustained even under denialism. Here, for example, is Waller on resentment: “Neither Matthew nor Donna is morally responsible, and neither justly deserves reward or punishment, but that fact – and my recognition of it – does not preclude reactive feelings of resentment and gratitude” (Waller 2011, p. 200). This approach, however, seems inadequate, for it is unclear how Waller has the resources to make such claims: what, for instance, does it mean to say that resentment is not “precluded” when, according to him, no one deserves to be resented? Similarly, why is being the focus of the resentment of others (when we assume that one is not morally responsible in any way for whatever one did to present the temptation of resentment) different from being unjustly punished?

Turning now to gratitude, Waller does not deny the importance of gratitude, but simply thinks that we can get it all without assuming moral responsibility. As he says, “Nonetheless, it seems to me that understanding the deeper causes of character and behavior does nothing to compromise either admiration or gratitude. All parties to this discussion are aware of the profound biological causes at work in producing a mother’s deep love and devotion to her children; but that causal knowledge in no way reduces or demeans the wonderful nature of that love, nor the gratitude one feels for the sacrifices and love and acts of kindness bestowed by a loving mother” (Waller 2017: 13 March).

This again seems highly implausible. Clearly there is a biological origin and some built-in biological basis for the efforts and sacrifices of mothers but there are, after all, huge actual

and undeniable differences among mothers in their commitment, devotion, efforts and sacrifices. Viewing such efforts and sacrifices without assuming choice-based and desert-generating moral responsibility makes things very different. We can see this by reflecting on the idea of ingratitude. Normally we would consider those ungrateful to others (who have been continuously devoted to them, sacrificed greatly on their behalf, and made large contributions; and assuming no large opposing faults), to be at fault. It is a fault, for not giving their beneficiaries their dues, i.e. what they deserve; to be thereby unfeeling, unresponsive, and unjust. And we can sympathize with the deep disappointment of (say) such a mother, whose lifelong devotion, effort, and contribution is not appreciated, who is not as it were really seen as a choosing, acting person, and who does not get the gratitude she feels she deserves.

But what can the denialist say in response? Not very much, it would seem. Under denialist views, the mother operated as she was molded, an unfolding of the given. Some mothers are neglectful and abusive, others are loving and devoted, even heroic; neither type is morally responsible for their performance, and neither is deserving of appreciation in the true, credit-giving sense. The central form of gratitude that is based upon appreciation of past and indeed present agency and effort becomes dubious. There is an *equality of value*, as it were, among all mothers – the admirable one is not to be particularly appreciated and thanked. And if gratitude is not deserved, then ingratitude can hardly be such a grave fault as we normally feel that it is. As elsewhere, with gratitude denialists are living with meager resources, and offering us a near-starvation diet. The connection between appreciation, gratitude, and certain central forms of love further shows the psychological and ethical poverty of denialism.

Other denialists have opted for a more revisionary approach and tried, typically in a more defensive mode, to recruit substitutes for the familiar beliefs, reactions, and practices based upon the belief in free will and moral responsibility. For example: “Instead of blaming people, the determinist might appeal to the practice of moral admonishment and encouragement. One might, for example, explain to an offender that what he did was wrong, and then encourage him to refrain from performing similar actions in the future” (Pereboom 2001, p. 325). One can also *protest* his intransience (Pereboom 2017b).

But for a genuine denialist, this cannot go very far. If the admonished offender does not respond adequately, even after we protest, she cannot, after all, be blamed. Since inducing a sense of responsibility-based blame is forbidden, people are unlikely to grow up with a dependable set of dispositions to behave responsibly – they will not even be acquainted with the notion, except as that held by other, mistaken people, who have been misled into believing in personal responsibility and desert. Denialists might still regard wrongdoing as an acceptable reason to weaken or dissolve a personal relationship (see Pereboom 2001, p. 201. Cf. Scanlon 2008). But this covers only a limited part of the spectrum of human interaction, as covered by the notion of blame.

Indeed, prominent denialists have wanted to specify certain senses of responsibility that will remain, even without free will, and these need to be distinguished from moral responsibility. As Waller puts it, “a world devoid of moral responsibility would not lack all individual responsibility; it would leave ample room for take-charge responsibility and increase the likelihood of exercising it well ... It enables us to exercise effective control, make our own decisions and choices, reflect carefully on what we deeply value, and manage our own lives” (Waller 2011, p. 278). Pereboom seems to be making a similar move when he speaks about “Moral responsibility without basic desert” (Pereboom 2014: Ch.6), which seeks to enable certain reformist notions of responsibility, obligation and blame without the reactive attitudes.

Here, of course absolute (traditional) or partial (i.e. compatibility pluralist) compatibilists will want to push back. They will say things such as that, if most adults most of the time can

“manage their own lives,” then why others cannot hold them accountable for their actions, expect them to behave responsibly and, at least sometimes and in some forms, blame, resent and otherwise sanction them when they do not? But this goes back to the disagreements on the compatibility question that we saw earlier on.

As we saw with regard to gratitude, denialists seek to combat claims that self-respect, appreciation, love and meaning in life, among other matters, are under severe threat, in two broad ways. They emphasize the ways in which the assumption of free will and moral responsibility is unnecessary; and try to offer substitutes that do not depend upon such ideas. These moves can go some distance. For example, clearly there are forms of love, such as loving one’s baby, that do not respond to free will at all. Even with gratitude, certain aspects such as feeling joy as a response to what one has done and expressing one’s joy can be retained even under denialist beliefs (Pereboom 2017c, p. 132). Much of human appreciation clearly is also not related to choice and agency, such as when we appreciate a person’s sense of humor or natural inborn singing talents; or, indeed, the beauty of nature. Sommers (2007) and Milam (2016) argue that we can live quite well without the reactive attitudes, and maintain valuable human relationships such as friendship (but see for example Shabo 2012).⁶

Yet it is difficult to reject the great dependency of forms of self-respect and respect for persons, or the appreciation of the efforts and sacrifices of others, on the ideas of free choice and moral responsibility. If everything that we and those we care about do is merely an unfolding of the given, which is ultimately beyond our control, then fundamental aspects of humanity such as those concerning the attaining of value, gratitude and appreciation-based love, are under grave threat. If not altogether senseless, they become much impoverished. And denialism of course gives up even on compatibilist understandings of these matters, not only libertarian ones.

The connection of these to the issue of meaning in life seems direct. My view is here as well largely pessimistic, seeing the free will problem as the graveyard of our noblest aspirations and deepest beliefs; and showing life to be profoundly absurd (Smilansky 2000: section 11.5, 2022). The implications of the denial of libertarian free will are enormous, and this is further enhanced if one also rejects compatibilist forms of free will and moral responsibility. For then, *all one’s attainments would ultimately not be to one’s credit, but merely an “unfolding of the given.”* This has obvious deflationary implications for the attainment and accumulation of meaning. Moreover, on my view, what this entails for our sense of self-respect, appreciation of others, and indeed the very sense of self is so momentous, that it leads to the *positive* need for demeaning forms of self-deception and illusion (as will be noted later), which further impacts meaning in life and the meaning of life.

Matters are of course complex, and there are more local advantages in denialist attitudes. One issue that has not received sufficient attention concerns the way individuals can be released from some burdens of the self, deal with overwhelming sense of guilt through the denial of their freedom and responsibility for things done or not done in the past, and other such potential “therapeutic” affects (see, for example, Spinoza 1955; Smilansky 2000: section 10.3; Strawson 2003).⁷

Beyond the pragmatic dangers (e.g. to motivation), human life risks losing a deep sense of value and of meaning that are intimately connected to the idea of free and responsible agency, and of what one acquires by the way in which one exercises them. Denialism is “the great eraser,” disconnecting human life from central aspects of a backward-looking sense of desert for one’s goodwill, efforts, and contributions; and these are crucial sources for generating (self)-respect, a sense of value and appreciation.

6 We Should Opt for Radical Change (Position 5)

The final of the five major positions making up denialism is that we should opt for radical change. In large measure, this follows from the combination of Positions 3 and 4, namely, the awfulness of the status quo and the viability of alternatives. Yet there is a need for an independent discussion of this claim. Compare robust egalitarianism: it is one thing to say that a certain state of affairs that radically differs from prevailing social reality in (say) a Western democracy is unjust, according to such egalitarianism, and quite another to say that we ought to make the radical transition to an alternative state. The risks and “transition costs” might be overwhelming – involving, for example, violent confiscation of private property, which only a dictatorship could achieve. There is further work that needs to be done, even if one feels that one can point out the faults of the present and sees a more attractive alternative. This is also the case with Position 5 of denialism.

One place to start would be a broad “Humean” tradition that is mostly due to P.F. Strawson’s highly influential “Freedom and Resentment” (Strawson 1962/2003). Interpreting Strawson is a matter of debate (see, e.g., McKenna and Russell, eds. 2008). For our purposes we will focus on his claim that philosophy and indeed reason are weak in the free will context, so that our beliefs and practices will in fact be closely intertwined with the “reactive attitudes,” attitudes such as resentment, indignation, and gratitude, which simply assume that most people nearly all of the time are proper targets for the attribution of responsibility for their choices and actions. This, Strawson claims, makes determinism, or indeed any other free will–related worry, rather moot, for human beings seem by and large hard-wired to hold themselves and others as morally responsible.

However, Strawson’s “no need to worry” claim is far too strong, and must be viewed as complacent (as I show in detail, Smilansky 2001; see also Sommers 2012). History and comparative studies of current societies show no indication of societies where people are not held responsible in any sense. This makes the mountain that optimistic denialists have to climb much higher; a point to which we shall come later on. Yet at the same time, these studies also clearly show that in no way can we rest assured that elementary moral distinctions and constraints related to the Core Conception will be respected. Humans have been and are all too ready to blame and punish collectively, disregard agency and innocence, not caring about ability (in any sense) to have avoided the given sanction, and the like.

Galen Strawson has put forward a strong case for the phenomenological inevitability of a sense of freedom and responsibility as one acts, as a constitutive part of agency (most recently in Strawson 2021). But for all its salience, this as well cannot safeguard the proper moral attitudes, from the possible loss of confidence and cynicism that an emerging belief in denialism can generate. History shows that even if our personal phenomenology has a “built-in” sense of libertarian free will and responsibility, this cannot safeguard our moral beliefs, attitudes, and practices – particularly in the face of consequentialist and similar sorts of pragmatic temptations.

Three concerns need to be taken up here at this stage, in the light of a more moderate stance on these issues. On the one hand, the relative “plasticity” of humanity is a basis for a more mitigated skepticism concerning denialism, in a way that connects, for example, to the doubts we saw above about the ability of denialism to safeguard deontological constraints. Humans are not limited in their interpersonal interactions and expectations as P.F. Strawson thought, nor is the phenomenology of personal experience emphasized by Galen Strawson

sufficient, and evil can ensue. A second point is that even if the extreme claim that no real change is possible is rejected, the centrality of basic human reactions and phenomenology surely remain important. Denialism is arguably too misleading, it does violence, to the connection between the intimate fabric of human life and assumptions about responsibility.

A third point is more favorable to denialism. It is that with some matters, such as the sense of the meaning of social practices and individual achievements, there is more leeway for the denialists than P.F. Strawson would allow. As we saw, the Core Conception and assumption of free will and moral responsibility are even today not assumed for all forms of love, or friendship, or sense of meaning in life. And beyond that, denialism can to some extent offer novel substitutes for the free will and responsibility-based attitudes and practices.

The thought is that the general focus on human emotions (as compared, say, to questions of innocence and justice) can serve denialists. This relates to a second major aspect of P.F. Strawson's thought, his emphasis on our emotions; which appears also, in different ways, with other philosophers, such as Galen Strawson (1986/2010); Honderich (1988); Double (1991); Wallace (1994); and Russell (1995) – the first three are denialists. The temptation is to bypass the cognitive difficulties, and find shelter in self-sustaining emotions. This can be taken by compatibilists to dismiss denialist worries (as with P.F. Strawson). But it seems to me that once we leave behind P.F. Strawson's assurance of built-in stability, such an approach cannot make much of what matters sufficiently safe, in practice, from free will-related doubts. As Bernard Williams reminded us, "Blame that is perceived as unjust often fails to have the desired result, and merely generates resentment" (Williams 1995, p. 15; see also Wolf 1981). A similar attempt by denialists to alleviate worries about the consequences of abandoning the common paradigmatic beliefs of the Core Conception should face similar difficulties. Our emotions in themselves cannot sustain a decent form of personal and communal life, in the face of the destruction of the sustaining beliefs. The difficulties of this non-cognitivist direction can be seen for example from the worries about gratitude we noted above.

Very little work has been done on evaluating the possibilities, risks, and dangers of a transition to a new denialist order, and the abandonment of the Core Conception-based Community of Responsibility. The grave worries we saw above, both of the dangers of trying to defend deontological constraints without belief in free will and moral responsibility, and the "softer" worries about personal life, self-respect, and meaning in life, lead to the reasonableness of skepticism about the feasibility and safety of trying to move to a radically different, denialist order.

6.1 *Concluding Reflections and the Needs for Further Research*

The traditional free will problem has encompassed the two traditional questions: roughly, that of the existence of libertarian free will, and the compatibility question of whether free will and moral responsibility can exist without LFW. Further questions which concern the goodness and the badness of the Core Conception-based status quo, the possibility for change, and the viability of alternatives and overall desirability of change, have all been recent additions to the free will debates. While great progress has been made, these topics remain greatly under-discussed, as compared to the two traditional questions.

Beyond the first two questions, denialism typically involves some ambiguity, largely as between being *reformist* versus *revolutionary*. Reformist denialists will suggest the need for change, but will typically see it as not being too deep or inclusive. Much in common life, they will claim, can remain more or less in its current form, even without belief in free will and moral responsibility. Most of our emotions, for example, can be either sustained or find similar

versions available even to denialists. In general, it makes sense for denialists to be revolutionary when in offensive mode, focusing on the deficiencies of their rivals, and reformist when defending against their rivals' claims that life without belief in free will and moral responsibility would be dangerous and bleak. And as we saw, this is indeed what we largely find, although there are of course differences in position (and partly in rhetoric, perhaps in temperament), among the denialists. One large project that denialism needs to address is *to clarify how revolutionary it is, and the extent to which this is a normative choice or a practical and pragmatic necessity*. The acceptability of denialism will be influenced by its answers to these questions.

A second important project for denialism is to consider the risks of the sought-after state of affairs, and what we called the "transition issues," that is, how one gets from the current, largely Core Conception-based status quo to the denialist new order. Who will be the agents of change? What are the prices and dangers? How are these to be avoided, or at least limited? Such questions have hardly been addressed. The opponents of denialism have an easier time here. Compatibilists, even reformist compatibilists such as Vargas (2013), have much more to work with in common beliefs, reactions, and practices. Those who tend to be wary of change such as myself (e.g. Smilansky 2000: Part 2, 2022) still need to give an account of how the worse features of the status quo might be dealt with (such as concerning punishment), but by and large maintain that no big changes can be risked, and should not be attempted.⁸

As we saw, the issues here combine in complex ways both questions of philosophical truth and plausibility, and very different ones of the pragmatic likelihood of success. It might, for example, be argued that there is more of a philosophical basis than has seemed to me to be for safeguarding deontological constraints, or perhaps even for viewing punishment as just, for denialists. Yet at the same time one could nevertheless doubt whether pragmatically these would suffice in the real world, if people would begin to think that there was no free will, moral responsibility, and desert. Similarly, one could argue that denialism philosophically harms much of the basis for meaning in life, so that if there is no free will and moral responsibility this is a dark truth indeed. Yet one could also hold that philosophy notwithstanding, most people would still, in fact, retain a sense that their lives were meaningful and, hence, pragmatically, applying denialism would not be so worrisome. The interaction between truth (including moral truth) and pragmatic concerns, in such contexts, is another topic that needs to be much more deeply discussed.

A third question for denialists is the extent to which they would like to remain with a radical, Absolute Denialism, or might want to opt for a milder, partial, Moderate Denialism, which accommodates some of the insights of compatibilism. This question can be addressed from within the traditional compatibility question but, as we saw, it matters a great deal when we come to address the newer questions and Positions 3–5 of denialism. Arguably, many of the aims of denialism for reform in punishment and in our reactive attitudes may be better served by accepting some of the weight of compatibilism, while insisting that compatibilist free will and moral responsibility are not enough and therefore the insights of denialism also need to be considered. The more radical form that contemporary denialism has taken may be more a matter of continuing confusion on the Assumption of Monism (the thought that one cannot combine some denialism with some compatibilism), the excitements of philosophical battle and social criticism, and a deep optimism, than something that is necessary in order to create philosophical and practical accommodation with the insights of denialism. For example, much of the actual way in which Pereboom engages with the world resembles revisionist compatibilism such as that of Vargas (Vargas 2013). Instead of positioning free will and moral responsibility impossibly high, Vargas proposes a forward-looking compatibilist model

that is in many ways similar to Pereboom's, but avoids the philosophical and ethical impoverishment that denialists are led to, by interpreting things in a skeptical denialist way. Recent denialist attempts to talk about those senses of responsibility that *can* be accommodated within a denialist world-view further move denialism closer towards compatibilism.

My own view here combines a (i) "Fundamental Dualism" or compatibility pluralism, which seeks to combine the partial insights of both denialism and compatibilism, with (ii) Illusionism concerning free will and moral responsibility, which seeks to address the dangers of living with the truth concerning free will; including the philosophical and pragmatic limitations of denialism that we have seen. This is not the place to consider these positions in detail (see, respectively, for the pluralism Smilansky 1993, 2000: ch.6, 2008, 2011a, 2012; Pinku 2012; and for Illusionism; Smilansky 2000: Part 2, 2012, 2022; for the oddness of what emerges, see, 2013b). In a nutshell, I claim that the human condition here is inherently pluralistic – we being creatures who typically have a large measure of local compatibilist control, who ought to be treated as responsible agents within a Community of Responsibility, who are allowed to live out the consequences of our choices – but we are at the same time determined beings, operating as we were molded, and this often generates severe injustice and great limitations in value and meaning. The interaction between these two contrastive aspects and the dangers of full awareness are then mitigated by illusions, whether they be of libertarian free will or of the sufficiency of compatibilism. The prospects for a Moderate Denialism and (if my concern is reasonable) the need to be conservative and circumspect about change in beliefs, attitudes and practices, offer themselves as alternatives to the more extreme, optimistic, and confident currently dominant varieties of denialism.

A fourth future project for denialism (and its opponents) is to engage with the issue of actual reforms in punishment, particularly in certain western European countries; and explore their real connection with beliefs about free will. Denialist discussions so far have been quite sketchy and far too parochial (i.e. North American), and have not investigated in detail the intriguing approaches to punishment of certain western European countries. It is not at all clear what is going on in (for example) Scandinavian models of punishment, and to what their apparent success can be attributed. This seems a promising path for denialism concerning punishment. Yet skeptics about optimistic-denialism-on-moral-responsibility will doubt, first, whether such options are really viable for far less homogeneous societies with a huge and violent criminal population, such as the US. Second, they will also doubt whether in fact the success of those European models does not depend on cultural traditions of restraint and the *internalization of moral responsibility*, that the rehabilitation processes build on and actually foster, rather than rejecting moral responsibility as the denialists recommend.

Finally, while one ought to be wary of easily translating empirical psychological studies about the role of free will beliefs into philosophical conclusions, there is clearly an important role for empirical testing of folk attitudes on our topics. This area is a fifth one for future engagement for denialists. Experimental philosophy studies of the compatibility question have yielded complex, contrastive and inconclusive results (see, e.g., Knobe and Nichols 2017). Much more work needs to be done, particularly in exploring the questions relevant to Positions 3–5 of denialism. Compatibilist interpretations of the Core Conception remain within this paradigm, but denialism is much more radical. This *ipso facto* raises acute questions about the possibility of change in the denialist direction, about the benefits, and about the risks and harms. In recent years, there has been a very large empirical body of literature supporting skepticism about the benefits of the prospect of living without belief in free will and moral responsibility.

This seems to support concern and conservatism rather than the prevalent optimistic denialism. A recent survey will serve to show the breadth of the conclusions:

It has been repeatedly shown that attenuating the belief in free will leads to various negative outcomes, such as an undermined sense of agency (being in control of one's actions) (Lynn et al. 2014), self-alienation (Seto and Hicks 2016) ... decrease in perceived meaningfulness in life (Crescioni et al. 2015), corruption of helpfulness and increase in aggression (Baumeister et al. 2009), weakening of the feeling of gratitude (MacKenzie et al. 2014), and reduction of cooperative behavior (Protzko et al. 2016). The belief in free will is also associated with ... stronger sense of self-efficacy, life satisfaction, meaning in life, gratitude, greater commitment in relationships (Crescioni et al. 2015), learning from emotional experiences (Stillman and Baumeister 2010), and beliefs in morality (Bergner and Ramon 2013). (Baliavsky 2020)

Denialist efforts to address such concerns are required. There is in any case reason to see an unfounded optimism as deeply ingrained and very widely prevalent in the free will debate (Smilansky 2010; for an exception see Russell 2013). Denialism has two distinct, but related, features, which makes denialist optimism particularly difficult. The first is the poverty of resources that, as we saw, denialists have available to them, the meager diet which they offer humanity concerning matters such as justice, moral depth, self-respect, gratitude and meaning in life. The second is the radical nature of the change that seems required, particularly if Absolute Denialism is taken up; the distance between common sense and the paradigmatic Core Conception, and the alternative denialist world.

Denialism has, in the last generation, become a real alternative in the free will debate as well as a social force; and a great deal of progress has been made in exploring it. The philosophical ambitiousness of denialism, and its enormous moral and practical significance, make further research imperative.⁹

Notes

- 1 Pereboom (2014) has put forward the influential idea of "basic desert" as the fundamental litmus test on the compatibility question. According to this idea, the agent would deserve to be blamed (or praised) just because she has freely performed the action, irrespective of consequentialist or contractualist considerations. This notion is too narrow for our purposes, as compared to the Core Conception, to represent the paradigm of free will and moral responsibility. It also seems too strict, in that various participants in the free will debates who acknowledge free will and moral responsibility probably do not share it.
- 2 The most effective belief for defending this Core Conception is (or would be, if it made sense and existed) robust libertarianism: the idea that although human beings are certainly affected and to some extent constrained by their heredity and environment, operating through their character, desires and beliefs, they nevertheless are able to *transcend* these influences. Hence when looking back one can see oneself as having been in large measure "where the buck stops," as a self-determining cause who was not completely determined to choose and act as one has. One could, strictly, have decided to have done otherwise, exactly as one was, and this option was actually within one's control. Robust LFW makes one morally responsible for one's choices and actions in a similarly robust sense, deserving of a whole set of moral and a-moral evaluations, reactions, and treatments.
- 3 There is of course criticism of punishment unrelated to the denialist critique; see for example Boonin (2008).
- 4 It might be argued that if there is no free will in any sense, we must abandon the Core Conception, the very notion of desert ceases to make any sense, and with it any sense of fairness or injustice based on it. And then perhaps the pragmatic justification of punishment (or of other attitudes and

practices) can proceed undisturbed. But this argument is unconvincing. Desert is a way of justifying divergence from a moral baseline. If desert becomes impossible, this does not mean that the moral baseline has somehow disappeared and that “anything goes” (cf. Smilansky 1996a, 1996b). If the moral baseline is that everyone ought to be treated as innocent unless proven guilty (through his or her free actions), then, if no one can become guilty, the moral baseline of innocence remains. Not to respect it would be unjust. It is thus a mistake to believe that since denialism rules out control-based desert, there is no justice (or injustice) in a morally deep denialist world.

- 5 Utilitarian-like consequentialists would give no moral weight to this question, for all that matters is the overall usefulness of punishment (e.g. Sidgwick 1907/1963; Smart 1961; Levy 2012).
- 6 Sommers started off as a denialist, but his more recent work, such as Sommers (2012), is meta-ethically relativistic. It is not clear how “denialist-friendly” such a view can be, but it is certainly not supportive of the Core Conception paradigm.
- 7 Paradoxically, belief in (the two first positions of) denialism also has the perverse moral advantage of potentially making one especially morally worthy, because one does not think that one’s moral worth is at stake in the free will–related ways. This is enhanced if one chooses morally, believes in denialism, but denialism is at least partly false and there is some free will–related moral worth. See Smilansky (1994), and a response in Double (2004); Smilansky (2000 section 10.1).
- 8 There might be technological changes that would change our options radically, but these lie sometime in the future, are largely unknown, and I will not consider this issue here.
- 9 I am very grateful to the editors for the invitation to write this essay; and to Joe Campbell, Iddo Landau, Sam Lebens, Arad Levin, Kristin Mickelson, Guy Pinku, Daniel Statman, Galen Strawson, and Alan White, for helpful comments on various drafts.

Bibliography

- Beliaevsky, V. (2020). *Freedom, Responsibility and Therapy*. London: Palgrave Macmillan.
- Berlin, I. (1980). From hope and fear set free. In: *Concepts and Categories*. Oxford: Oxford University Press.
- Boonin, D. (2008). *The Problem of Punishment*. New York: Cambridge University Press.
- Bradley, F.H. (1927). *Ethical Studies*. Oxford: Clarendon Press.
- Caruso, G.D. (2016). Free will skepticism and criminal behavior: A public health-quarantine model. *Southwest Philosophy Review* 32: 25–48.
- Corrado, M.L. (2017). Insanity and free will: The humanitarian argument for abolition. In: *The Insanity Defense: Multidisciplinary Views on Its History, Trends, and Controversies* (ed. M.D. White). Santa Barbara, CA: Praeger.
- Corrado, M.L. (2018). Free will, punishment, and the burden of proof. *Criminal Justice Ethics* 37: 55–71.
- Double, R. (1991). *The Non-reality of Free Will*. New York: Oxford University Press.
- Double, R. (2004). The ethical advantages of free will subjectivism. *Philosophy and Phenomenological Research* 69: 411–422.
- Feldman, F. (2004). Freedom and contextualism. In: *Freedom and Determinism* (ed. J.K. Campbell, M. O’Rourke and D. Shier). Cambridge, Mass.: MIT Press.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control*. Cambridge: Cambridge University Press.
- Frankfurt, H.G. (1988). Freedom of the will and the concept of a person. In: *The Importance of What We Care About*. Cambridge: Cambridge University Press.
- Glover, J. (1983). Self-creation. *Proceedings of the British Academy* 69: 445–471.
- Hart, H.L.A. (1970). *Punishment and Responsibility*. Oxford: Clarendon Press.
- Hawthorne, J. (2001). Freedom in context. *Philosophical Studies* 104: 63–79.
- Honderich, T. (1988). *A Theory of Determinism*. Oxford: Clarendon Press.
- James, W. (1896/1956). The dilemma of determinism. In: *The Will to Believe and Other Essays in Popular Philosophy*. New York: Dover Publications.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.
- Klein, M. (1990). *Determinism, Blameworthiness and Deprivation*. Oxford: Oxford University Press.

- Knobe, J. and Nichols, S. (2017). Experimental philosophy. *Standard Encyclopedia of Philosophy*.
- Lemos, J. (2016). Moral concerns about responsibility denial and the quarantine of violent criminals. *Law and Philosophy* 35: 461–483.
- Levy, N. (2011). *Hard Luck*. Oxford: Oxford University Press.
- Levy, N. (2012). Skepticism and sanction. *Law and Philosophy* 31: 477–493.
- McKenna, M. and Russell, P. (eds) (2008). *Free Will and Reactive Attitudes: Perspectives on P.F. Strawson's "Freedom and Resentment"*. Farnham, UK: Ashgate.
- Milam, P.-E. (2016). Reactive-attitudes and personal relationships. *Canadian Journal of Philosophy* 46: 102–122.
- Morris, H. (1976). Persons and punishment. In: *On Guilt and Innocence*. Berkeley and Los Angeles: University of California Press.
- Murphy, J.G. (1973). Criminal punishment and psychiatric fallacies. In: *Punishment and Rehabilitation* (ed. J.G. Murphy). Belmont: Wadsworth.
- Nagel, T. (1986). *The View From Nowhere*. Oxford: Oxford University Press.
- Nagel, T. (2012). Moral luck. In: *Mortal Questions*. Cambridge: Cambridge University Press.
- Nelkin, D.K. (2011). *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.
- Parfit, D. (2011). *On What Matters*. Oxford: Oxford University Press.
- Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2014). *Free Will, Agency and Meaning in Life*. New York: Oxford University Press.
- Pereboom, D. (2017a). A defense of free will skepticism: Replies to commentaries by Victor Tadros, Saul Smilansky, Michael McKenna, and Alfred R. Mele on *Free Will, Agency, and Meaning in Life*. *Criminal Law and Philosophy* 11: 617–636.
- Pereboom, D. (2017b). Responsibility, regret and protest. *Oxford Studies in Agency and Responsibility* 4: 121–140.
- Pereboom, D. (2017c). Skeptical views about free will. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy). New York: Routledge.
- Pinku, G. (2012). Morally embedded selves and embedded compatibilism. *Philosophica* 85: 67–89.
- Russell, P. (1995). *Freedom and Moral Sentiments*. New York: Oxford University Press.
- Russell, P. (2013). The philosophy of free will. In: *Compatibilist-fatalism: Finitude, Pessimism, and the Limits of Free Will* (ed. P. Russell and O. Deery). Oxford: Oxford University Press.
- Scanlon, T.M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.
- Shabo, S. (2012). Where love and resentment meet: Strawson's intrapersonal defense of compatibilism. *Philosophical Review* 121: 95–124.
- Sidgwick, H. (1907/1963). *The Methods of Ethics*, 7e. London: Macmillan.
- Smart, J.J.C. (1961). Free will, praise and blame. *Mind* 70: 291–306.
- Smilansky, S. (1990). Utilitarianism and the 'punishment' of the innocent: The general problem. *Analysis* 50: 256–261.
- Smilansky, S. (1993). Does the free will debate rest on a mistake? *Philosophical Papers* 22: 173–188.
- Smilansky, S. (1994). The ethical advantages of hard determinism. *Philosophy and Phenomenological Research* 54: 355–363.
- Smilansky, S. (1996a). Responsibility and desert: defending the connection. *Mind* 105: 157–163.
- Smilansky, S. (1996b). The connection between responsibility and desert: the crucial distinction. *Mind* 105: 385–386.
- Smilansky, S. (2000). *Free Will and Illusion*. Oxford: Oxford University Press.
- Smilansky, S. (2001). Free will: From nature to illusion. *Proceedings of the Aristotelian Society* 101: 71–95.
- Smilansky, S. (2003). Compatibilism: The argument from shallowness. *Philosophical Studies* 115: 257–282.
- Smilansky, S. (2005). Free will and respect for persons. *Midwest Studies in Philosophy* 29: 248–261.
- Smilansky, S. (2008). Free will and fairness. In: *Essays on Free Will and Moral Responsibility* (ed. N. Trakakis and D. Cohen). Cambridge: Scholars Publishing.

- Smilansky, S. (2010). Free will: Some bad news. In: *Action, Ethics and Responsibility* (ed. J.K. Campbell, M. O'Rourke and H.S. Silverstein). Cambridge, Mass.: MIT Press.
- Smilansky, S. (2011a). Free will, fundamental dualism and the centrality of illusion. In: *The Oxford Handbook of Free Will* (ed. R. Kane), 425–441. New York: Oxford University Press.
- Smilansky, S. (2011b). Hard determinism and punishment: a practical *reductio*. *Law and Philosophy* 30: 353–367.
- Smilansky, S. (2012). Free will and moral responsibility: The Trap, the appreciation of agency, and the bubble. *Journal of Ethics* 16: 211–239.
- Smilansky, S. (2013b). Free will as a case of “Crazy ethics”. In: *Exploring the Illusion of Free Will and Moral Responsibility* (ed. G.D. Caruso). Lanham, MD: Lexington Books.
- Smilansky, S. (2016). Parfit on free will, desert, and the fairness of punishment. *Journal of Ethics* 20: 139–148.
- Smilansky, S. (2017a). The free will problem: nonstandard views. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy). New York: Routledge.
- Smilansky, S. (2017b). Pereboom on punishment - funishment, innocence, motivation, and other difficulties. *Criminal Law and Philosophy* 11: 591–603.
- Smilansky, S. (2019). Free will denial and deontological constraints. In: *Free Will Skepticism in Law and Society* (ed. E. Shaw, D. Pereboom and G.D. Caruso). New York: Cambridge University Press.
- Smilansky, S. (2022). Illusionism. In: *The Oxford Handbook of Moral Responsibility* (ed. D. Pereboom and D. Nelkin). New York: Oxford University Press.
- Sommers, T. (2007). The objective attitude. *Philosophical Quarterly* 57: 321–341.
- Sommers, T. (2012). *Relative Justice*. Princeton: Princeton University Press.
- Spinoza, B. (1955). *Ethics*. New York: Hafner.
- Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies* 75: 5–24.
- Strawson, G. (1986/2010). *Freedom and Belief*. Oxford: Oxford University Press.
- Strawson, G. (2003). “You cannot make yourself the way you are.” A Conversation with Tamler Sommers. *Believer* (online).
- Strawson, G. (2021). The phenomenology of free agency. In: *Routledge Handbook of the Phenomenology of Agency* (ed. C. Erhard and T. Keiling). Abingdon: Routledge.
- Strawson, P.F. (1962/2003). Freedom and resentment. In: *Free Will* (ed. G. Watson). Oxford: Oxford University Press.
- Vargas, M. (2013). *Building Better Beings*. Oxford: Oxford University Press.
- Vilhauer, B. (2013). Persons, punishment and free will skepticism. *Philosophical Studies* 162: 143–163.
- Wallace, R.J. (1994). *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.
- Waller, B. (2011). *Against Moral Responsibility*. Cambridge, Mass.: MIT Press.
- Waller, B. (2015). *The Stubborn System of Moral Responsibility*. Cambridge, Mass.: MIT Press.
- Waller, B. (2017). *Syndicate; Symposium on the Stubborn System of Moral Responsibility*.
- Waller, B. (2018). *The Injustice of Punishment*. New York: Routledge.
- Williams, B. (1995). How free does the will need to be? In: *Making Sense of Humanity*. Cambridge: Cambridge University Press.
- Wolf, S. (1981). The importance of free will. *Mind* 90: 386–405.

Revisionism

Manuel R. Vargas

1 Introduction

A theory of x is revisionist if the truth of the theory's account of x is in conflict with commonsense views about that thing. Those commonsense views that are repudiated by the revisionist theory may include views about x 's theoretical commitments, the nature and status of practices involving x , and the characteristic propositional attitudes and inferences involving x . In its canonical forms, revisionist theories are *preservationist*, i.e., accounts that seek to retain the term, concept, or practice for which revision is proposed. Revisionist theories of some phenomenon x typically contrast with *eliminativist* accounts, which seek to expunge talk, thinking, or practices of x . However, revisionist accounts also conflict with *conventional* accounts, where neither elimination nor revision is called for. Revisionist theories have been proposed for a wide range of things, including race, personal identity, propositional attitudes, and various notions within the philosophy of science (for discussion see Nichols 2015, pp. 56–62 and Vargas 2013a, pp. 73–75).

A revisionist theory of *free will* is one that holds that the correct account of free will is, in some or another way, at odds with our ordinary understanding of free will's nature, concept, or associated practices. It typically conflicts with eliminativism (Strawson 1986; Pereboom 2001; Levy 2011; Caruso 2015), in maintaining that we can and should preserve characteristic talk, judgments, and practices bound up with free will. It departs from conventional theories in maintaining that a philosophically satisfactory account of free will is in conflict with ordinary understandings, theories, or practices concerned with free will.

Typically, there is a *methodological* component to revisionism; the revisionist theory is the ensuing distinctive *substantive* results that come about from adopting the methodology. Methodologically, revisionism depends on distinguishing between what we – the folk, or competent language users, or ordinary practitioners – think about something from what, all things considered, we *ought* to think about that thing. Substantively, a theory is revisionist to the extent that the substance of the all-things-considered proposal conflicts with some aspect of folk or “commonsense” thinking. A theory is not revisionist if, for example, the theorist has mischaracterized folk views of x and offered an all-things-considered theory of x that conflicts

with that mischaracterization. Such an account would not satisfy the requirement of being in conflict with actual folk views. Similarly, a view would fail to be revisionist if the prescriptive (or all-things-considered) proposal involved resolving or refining commitments within the constraints of existing indeterminacies in folk thinking. Such “revisions” are generally too trivial to count as the basis of a distinctive theoretical or methodological option.

The scope of a proposed revision can vary according to a given account’s assessment of the relationship of free will to other notions, and the basis of those other notions. Revisionism about conceptual commitments, for example, might not entail revisionism about existing social practices. Or, it might only entail revisions about those practices in very tightly circumscribed conditions. An account might be revisionist about the meaning of a concept, without being revisionist about how the concept is to be deployed in everyday usage. This sort of revisionism might insist that the account preserves paradigmatic inferences about free will, apart from those concerned with its nature. Depending on what one is a revisionist about (e.g., concepts, attitudes, practices) and the degree of one’s revisionism, one’s revisionism may be a more or less salient feature of the account.

Revisionism about free will has been motivated by a variety of worries regarding putative elements of received commitments about the concept, meaning, or nature of free will and morally responsible agency. For example, revisionists have argued that folk conceptions of *agency* are metaphysically implausible (Doris 2015a; Nichols 2015), that ordinary conceptual and metaphysical commitments about *freedom* and *responsibility* are implausible (Vargas 2004b, 2013a), that considerations about phenomenology (Deery 2015), reference (Heller 1996; McCormick 2013, 2015), and natural kinds (Deery, in progress) favor a form of revisionism, and that eliminativist views neglect a more moderate and plausible alternative than the strong forms of agency demanded by some accounts of free will (Hurley 2000; Vargas 2005a).

The basis of doubts about received or pre-revised notions of agency, freedom, and responsibility has also varied. To make the argument that ordinary or folk views of free will are untenable, philosophers have drawn from various sources, including: experimental data concerning folk notions (Vargas 2013a; Nichols 2015); considerations from the history of ideas (Nichols 2007b) and culture and religion (Vargas 2013a, 2016); reflections on phenomenology and language (Deery 2015; McCormick 2016); the presumptions of moral concepts (Smart 1961; Arneson 2003), and reflections on the fit between folk notions and findings in cognitive science and experimental science (Doris 2002, 2015a; Nichols and Knobe 2007; Knobe and Doris 2010; Vargas 2013c, 2017; Weigel 2013; Nichols 2015; Faucher 2016).

Perhaps the most common form of revisionism about free will reflects a concern in the compatibility debate, claiming that the philosophically preferred view is one that conflicts with standing folk convictions about free will. One might allow that folk thinking has incompatibilist strands, in whole or in part, but prescribe compatibilism as a solution. To be sure, one can be a revisionist about free will without being committed to a specifically compatibilist revisionism. For example, a revisionist theorist might hold that everyday thinking about free will is committed to *causa sui* agency (Strawson 1986), but that our best philosophical theory supports a form of event-causal libertarianism as a suitable revisionist replacement. That said, it is unclear the extent to which some or any libertarians currently conceive of their accounts as revisionist (Cf. Vargas 2005a, p. 418, 424 n.8).

Revisionism about free will need not be grounded in aspects of the compatibility debate. For example, some accounts of free will maintain that the most appealing conception of it will not be sufficient to support retributive elements in our ordinary practices, and these theorists are often

prepared to acknowledge that they are potentially revisionists on this point (Wallace 1994, p. 228; Scanlon 1998, p. 275; see also the discussion in Nelkin 2011, pp. 175–177). Alternately, one might isolate grounds for revisionism in what we think about agency more generally. For example, one might hold that free and responsible agency exists but that it does not involve, for example, the psychological features we ordinarily take to ground it (Cf. Doris 2015a).

Despite its distinguished pedigree in the philosophy of mind, philosophy of race, and in the philosophy of science, it is only in recent years that revisionism has been regarded as a distinctive family of views in debates about free will and moral responsibility (Cf. McCormick 2016). Important precursors to self-described revisionist theories of free will include accounts by Smart (1961), Bennett (1980), Heller (1996), Singer (2002), Hurley (2000), and Arneson (2003). Among contemporary theorists, philosophers who have identified their proposals as revisionist in some or another way, including Walter (2004), Vargas (2004a, 2005a, 2013a), Nichols (2015), Weigel (2013), McCormick (2013, 2015), Doris (2015a). Others have explicitly (as we saw above with Wallace, Scanlon, and Nelkin) or tacitly (McGeer 2015) signaled an openness to their accounts being construed (in part) as a form of revisionism about free will and/or moral responsibility.¹

In the context of debates about free will, revisionism's appeal seems to be grounded in its putative resources for (1) explaining disagreements about conflicting intuitions and the persistence of debates about free will, and (2) re-anchoring our understanding of the epistemic, metaphysical, and normative basis for free will (and/or moral responsibility). Neither (1) nor (2) is uniquely claimed by revisionist theories. However, those who are drawn to the more ambitious forms of revisionist accounts pursue them partly because they believe that they enable one to acknowledge some powerful aspects of everyday thinking (and, thus, to remain in conversation with those gripped by these convictions) without thereby making their positive or prescriptive accounts of free will beholden to folk convictions. To be sure, whether the rehabilitation of common sense seems required, interesting, and/or promising depends a great deal on whether one is inclined to think there is something infelicitous about ordinary convictions about free will. Conventional compatibilists and libertarians tend not to.

In what follows, I will treat *free will* as a placeholder term, without committing to any particular view of how it is to be understood. However, depending on how one understands the various involved ideas, revisionism about free will may entail revisionism about moral responsibility, responsible agency, deliberation, and various notions of moral desert. Given this fact of the state of play in contemporary philosophical discussions, I will sometimes move between talk of revisionism about free will, and talk of revisionism about responsible agency and moral responsibility. This is not ended to reflect an endorsement of a substantive view about how to characterize free will. Rather, it is only an effort to capture the range of philosophical approaches that might plausibly be construed as revisionist about free will.²

The structure of the rest of this chapter is as follows: First, I discuss some of the methodological foundations of revisionist accounts. Then, I discuss various grounds that have been offered for revisionism. In a subsequent section, I consider some objections directed at recent revisionist developments. Finally, I conclude with thoughts on the direction of revisionist theorizing.

2 Methodological Foundations

It is undoubtedly true that within the literature on free will, methodological approaches have varied considerably (Double 1996). However, it is also certainly true that a concern for ordinary convictions about freedom, and for grounding philosophical arguments in those convictions, has been a widespread feature of philosophical work on free will (Nahmias

et al. 2006; Nichols 2006, 2007a; Nahmias 2011; Björnsson and Persson 2013). A robust philosophical preoccupation with “intuitiveness” and coherence with folk commitments in the context of free will might explain why interest in convictions about free will were an especially prominent feature of the first wave of experimental philosophy and empirically informed approaches to moral psychology (Nahmias 2006; Doris, Knobe, and Woolfolk 2007; Nahmias, Coates, and Kvaran 2007; Roskies and Nichols 2008).

A wide range of philosophical methodologies maintain that an adequate philosophical account of many, perhaps most, phenomena begins from the everyday meaning, usage, or thinking about the particular phenomenon in question. More ambitiously (or perhaps its reverse?) some have thought that the elucidation of some “ordinary” or “common sense” concepts *just is* the task of metaphysics, and sometimes, philosophy more generally (Strawson 1959; Lewis 1973; Jackson 1998). Others have thought that the proper task of philosophy should seek to explain how widely accepted ideas could be true (Nozick 1981). Coherence with folk commitments is a crucial, even constitutive desideratum of these philosophical methodologies. Given the history of debates about free will in the analytic traditions – including debates about the meaning of “can” in the 1950s and 1960s, and the metaphysics of ability in the 1960s and 1970s (Vargas 2011) – it is perhaps unsurprising that a good deal of the literature has been structured by sometimes explicit but oftentimes tacit presumptions about the importance of folk thinking about free will. Against this background, revisionism represents a methodological break in the consensus about the role of folk commitments for philosophical theory-building.

Revisionist methodologies usually start by distinguishing at least two questions. First, what is the status of commonsense thinking about the thing in question? Call this the *diagnostic question*. Second, what sort of theory should we have about that thing, all things considered? Call this the *prescriptive question*. The disjunctive nature of the diagnostic question – what is the nature and plausibility of ordinary commitments about *x*? – can be further separated into two component questions. Shaun Nichols (2015) helpfully distinguishes between a *descriptive* question and *substantive* question within the broader diagnostic question. The descriptive question concerns our (typically folk, but potentially philosophical) beliefs in some domain. The substantive question concerns the status of those beliefs – as true, false, implausible, and so on. What is helpful about subdividing the diagnostic question into two further questions, descriptive and substantive, is that it is one thing to identify folk commitments, it is another to decide that they are implausible or in need of revision. In practice, revisionists accounts tend to proceed by first identifying some problematic aspect of common sense and/or folk notions (of freedom, race, propositional attitudes, etc.), and then moving to a positive proposal for theorists (and perhaps the folk) to go forward without the problematic aspect.³

At this point, some might be inclined to protest that the purported distinctiveness of revisionist theorizing is hardly distinctive at all. Many philosophers accept something like the method of reflective equilibrium, on which theoretical revisions of considered beliefs and principles are permissible. Even on conceptual analysis-style accounts of philosophical methodology, revisions that depart from aspects of common sense or folk beliefs can be a relatively pedestrian feature of conceptual analysis, where the stakes can be something “suitable close to our ordinary conception” (Jackson 1998, p. 31). Moreover, relatively radical revisions can be licensed under particular kinds of conditions. For example, some are prepared to allow for the possibility of infinite concrete possible worlds on the grounds that this sort of counter-intuitiveness enables a reduction in the number of basic entities posited by the theory (Lewis 1973, p. 87). Thus, isn’t revisionism just fussiness about something that is already methodologically a commonplace?

When the alternative is confusion, fussiness can be a necessary virtue. Revisionist accounts are often motivated by the thought that a failure to be cautious about whether and when one is

intending to revise away from problematic folk notions has led to cross-talk and methodological confusions about what accounts of free will are endeavoring to provide (Vargas 2011; Nichols 2015). The thought seems to be this: in domains where folk beliefs, convictions, or practices are taken to have special weight in constructing a theory of that domain (e.g., moral and social categories), revisionist theories face special burdens licensing departures from folk thinking. Thus, it can be particularly useful to identify that the theory is revisionist, because it will have those special burdens. Relatedly, there are certain intuitions or judgments that the theory may not be in the business of trying to preserve or capture. So, marking the theory as revisionist is a way of being cautious about what is being revised and what is not thought to be in need of revision.

In short, failure to mark such differences is less a matter of declining to be fussy than it is a failure of intellectual hygiene. To leave the revisions unmarked runs the risk of presenting revisionist proposals as descriptions of folk commitments, and folk commitments as (tacitly and adequately) grounded prescriptions when they are not. (What is “suitably close to our ordinary conception” to count as our ordinary conception, anyway?) Conceptual analysis, philosophical explanation, and the method of reflective equilibrium too often fail to mark out what is putatively a part of everyday thinking and what is the product of revisionary philosophical labor. In failing to mark such distinctions, we risk running together ordinary convictions with philosophical innovation, and we leave ourselves open to unnoticed changes in topic. Distinguishing revisionist and conventional proposals clarifies the stakes and the options.

None of this is to deny that there are phenomena where conflicts with folk commitments are rightly regarded as an entirely unremarkable matter of course. In such cases – contemporary physics is a particularly clear example – there is little reason to identify the account as revisionist. However, in domains where accounts are not always understood to be revisionist, and where there is a heavy reliance on folk commitments in the construction of theories, there is special reason to be explicit about whether and when one is being revisionist. Thus, although revisionist theorizing can be found in a wide range of theoretical domains, its salience, degree, and significance vary by the context of our theorizing about those things.

One can be a revisionist about the meaning of terms, about the metaphysics associated with a term, about the concept, about judgments and inferences connected to the term, about practices associated with it, and even about emotional attitudes characteristic of the term and its associated practices. So, it would be useful to have a neutral term covering all these cases. Unless there is some matter where more precision is useful, in what follows I will generally speak of revisionism about commitments, intending this to cover the panoply of more particular varieties of revisionism. I will also put to one side some potential complexities concerning unintentional revisions, or revisions that are products of inconsistency in theoretical design. The general form of revisionism that is of interest in what follows is intentional revisionism away from some plausible understanding of current practices, beliefs, dispositions of judgment and attitudes – i.e., our commitments.

Some early accounts of free will revisionism (e.g., Vargas 2005a) failed to distinguish what we might call the *existence question* from the prescriptive question. The existence question concerns whether or not the considered property exists, or is instantiated. Different answers to the existence question are compatible with a revisionist prescriptive proposal. To see why, consider that one might hold that the best prescriptive account of x is revisionist, but that even after revision, that we should conclude that it does not exist.

The idea of true love might serve as an example. Juan might believe that true love is widely regarded as a monogamous tie between two souls that are perfectly fitted to one another. However, he might be inclined to accept a revisionist theory of true love, according to which true love is compatible with both the non-existence of souls and non-monogamous romantic relationships. For example, Juan might accept a threshold theory of love, according to which crossing

some threshold of mutual fulfillment and affection constitutes true love. Given the foregoing, it would be available to him to discover that he is a polyamorist materialist who has more than one true love. He might accept this revisionist theory for years without difficulty. But suppose Juan becomes a divorce lawyer and after a decade of practice he becomes jaded about the possibility of true love. If so, he might eventually conclude that true love, even on his revisionist account of it, does not exist. Perhaps some further fact – something discovered in the accumulating emotional debris of divorce proceedings – leads him to think that no one ever attains the threshold of mutual fulfillment and affection he took to be (his revised) notion of true love. So, Juan might be a revisionist about true love even while accepting that true love does not exist.

To be sure, most prescriptive proposals are offered on the presumption that the target property exists. In many contexts, there would be little that is theoretically appealing about a proposal to modify our thinking or practices concerned with *X*, if one did not think the proposed modification succeeded in tracking a viable (that is, existing or potentially existing) property in the world.

3 Arguments for Revisionism

In this section, I canvas some of the arguments for revisionism that have appeared in the recent philosophical literature. These include arguments from two-dimensional semantics, a debunking theory of the psychological mechanisms involved in incompatibilist convictions, doubts about the scientific plausibility of ordinary conceptions of agency, and various combinations of these views.

One motivation for incompatibilist pictures of free will is the idea that ordinary phenomenal experience has a libertarian character. That is, in ordinary deliberation about what to do, the phenomenal character of such deliberation is best understood as having libertarian commitments. Some have gone on to argue that this phenomenal character is reference-fixing (Caruso 2015). However, drawing from work by David Chalmers on two-dimensional semantics, Deery (2015) argues that phenomenal experience may well be libertarian, but that it also comes with “imperfect content” that provides veridicality conditions that appeal to whatever it is that normally causes those libertarian experiences. The punch line is that there is good reason to think *those* properties – i.e., those properties that ordinarily cause libertarian experiences – are in fact compatibilist-friendly properties, unthreatened by determinism. The result is a kind of error theory about folk phenomenology – it presents as libertarian in a way that cannot be sustained. However, it is also a kind of success theory about the properties that produce the error-ridden phenomenology. Deery goes on to recommend that our theories track this latter property, even though it conflicts with the evident phenomenology of choice-making.

Nichols (2015) maintains that the problem of free will arises because of some peculiarities of the psychology of explanation in the context of human action. On Nichols’s account, we have “an explanatory compulsion,” according to which we believe every action or event must have a cause. This compulsion to explanation lends itself to seeing human action as a causal (deterministic, in particular) product of prior events. However, in the case of human action the explanation yields a distinctive result. In first-personal cases of action explanation, when I consider why I acted the way I did, there is no immediately identifiable causal origin for my action. Nichols holds that when we cannot identify the causal sources of our own actions, we take this as evidence that our actions lack deterministic causes (42–50). The result is the genesis of the free will problem, namely, a conflict between our normal deterministic picture of explanation and a robust sense of our own action being indeterministic, i.e., the “indeterministic intuition.” However, our introspection is a flawed guide to the causal factors involved in our decision making, and, thus, “our intuition of indeterminism counts for nothing” (53).

Vargas (2013a) argues that folk commitments about free will and moral responsibility are not compatibilist or incompatibilist in a uniform way. Instead, there is oftentimes diversity across and within individuals about these commitments. Moreover, various rational, deliberative, and cognitive pressures regiment or systematize those commitments in ways that support compatibilist and incompatibilist convictions in the philosophical domain (2013a, pp. 49–51). However, the persistence of the free will problem, as a philosophical problem readily available to ordinary reflection, arises in large part because of the pervasiveness of specifically libertarian-friendly commitments (alternative possibilities, and perhaps a kind of sourcehood) that are supported by distinctive psychological and cultural scaffolding (2013a, pp. 40–43, 2016). These commitments are not necessarily had by everyone. Nevertheless, various aspects of human deliberation, explanation, and cultural narratives make those commitments appealing to a wide range of people. Evidence of the widespread presence of those libertarian commitments (about the basis of responsibility, about the nature of agency, and about the phenomenology of deliberation) can be found in experimental work (Nichols and Knobe 2007; Weigel 2013), broadly sociological data (Collins 1998; Nichols 2007b; Bourget and Chalmers 2014; Vargas 2016), and in the readiness with which people understand talk of abilities in the context of philosophical arguments for incompatibilism.

Gunnar Björnsson and Karl Persson (2012, 2013) have developed an account of the explanatory framework of human action according to which the widespread presence of incompatibilist commitments in the folk, at least with respect to talk about responsibility, is a byproduct of shifts between explanatory frameworks. As they see it, the ordinary concept of moral responsibility is bound up with our interests in identifying relevant motivational structures that play a suitable systematic role in action explanation. Thus, the ordinary concept of responsibility has been “shaped by conscious and unconscious interests and concerns that govern our practices of holding people responsible” (2013, p. 328). However, broadly incompatibilist regress arguments (e.g., Strawson 1994; Pereboom 2001, pp. 112–117) and appeals to determinism get their force from shifting attention away from the explanatory concerns that are salient for responsibility attributions.

Björnsson and Persson (2012) outline different possible readings of their account, including interpretations on which it is an account of the contextualist semantics of responsibility, an account on which it yields a story about our judgments of responsibility but not the (realist) nature of responsibility, and an account of some dispositions structuring expressivist views of responsibility. As they note, issues here are delicate. A variety of philosophical commitments bear on how we should understand the upshots of their approach, including: whether we have an independent account of the reference of moral terms, or moral responsibility in particular; accounts of the function of the involved concepts and practices of holding responsible; and the normative preconditions for holding people responsible (346–348).

Perhaps the most natural reading of Björnsson and Persson’s model is that it suggests an error theory for incompatibilist convictions, and correspondingly, a picture according to which the referent of talk about responsibility picks out some property compatible with a deterministic picture of the world. On their account, humans are prone to adopting faulty incompatibilist conclusions, and that absent some special remediation for a natural cognitive tendency, there will always be people who erroneously come to accept incompatibilist commitments about moral responsibility.

If that is how it is read, Björnsson and Persson’s picture suggests a form of connotational revisionism (Vargas 2013a, p. 88), according to which terms like “free will” and “moral responsibility” successfully refer, but for which we (pre-revision) had a variety of false beliefs about the nature of the thing referred to. However, the Björnsson and Persson model also admits of another possibility, according to which the tension in explanatory frames fails to

provide a determinate referent for talk of moral responsibility. On this account, the picture might favor a kind of denotational revision, that is, where there is a change in referent.⁴ Björnsson and Persson take no stand on these matters, but their proposal suggests a powerful and principled basis for adopting revisionism about moral responsibility.

In recent work, John Doris (2015a) has adopted a revisionist account of responsible agency, but on grounds disconnected from the traditional compatibility debate.⁵ Doris holds that ordinary thought and many strands of philosophical theorizing about agency are structured by a background commitment to view that the exercise of human agency consists in judgment and behavior ordered by self-conscious (and accurate) reflection about what to think and do. However, a substantial body of experimental considerations suggest that the core of this view – something Doris calls *reflectivism* – is false. He holds that our responsibility practices can and should be retained – they are “felicitous,” as he says – but that an accurate account of their basis involves a radically different understanding of the nature of human agency and the way in which responsibility practice interacts with responsible agency.

Although Doris is a kind of pluralist about responsible agency – he allows it may be attained in a variety of ways – he also argues that there is a relatively radical alternative conception of our agency that we should accept, something he characterizes as *collaborativism*. Collaborativism is the idea that optimal human reasoning is facilitated by social interaction (103). As Doris sees it, the process of reasons-giving, and various intra- and interpersonal explanations of action and statements of values, function less as an accurate statement of the springs of action than as a kind of social scaffolding that serves to coordinate future action and publicly establish commitments to particular standards of action evaluation. This is not intended as a characterization of how we understand ourselves, but rather, a characterization of how we ought to understand ourselves, given an empirically informed conception of our agency. Doris characterizes his view as conservative about the practice of moral responsibility but revisionary about the theoretical underpinnings (158). However, his revisionism is tempered by the thought that it is exceedingly difficult to demarcate folk theories about agency (i.e., whether and to what degree an account is revisionist may be a fundamentally indeterminate matter) (Cf. Doris 2015b).

4 Doubts about Revisionist Approaches

Most explicitly revisionist accounts of free will have proposed prescriptive theories of free will that are compatible with the truth of the thesis of determinism. Moreover, one can convert nearly any conventional compatibilist into a revisionist account simply by acknowledging that folk thinking has incompatibilist commitments, and that one’s compatibilist proposal is intended as a correction of defective folk thinking. So, one might wonder, is this evidence that revisionism is simply compatibilism by a different name?

In reply: no. As noted at the outset, what makes a theory revisionist is that it is in conflict with commonsense commitments. Revisionism is silent on the content of those commitments, and one can be a revisionist and offer an incompatibilist prescriptive theory. So, revisionism is not just compatibilism. It is true that conventional compatibilist theories of free will can be readily converted to revisionist proposals. However, such proposals come with distinctive costs. For example, the pressure to identify the grounds for making the revision becomes unavoidable for any newly revisionist theory, as does the explanation of why the proposed revision does not amount to a changing of the topic. If the aspirations of a theory change, what the theory has to show changes.

A critic might reply that everyone is, to some extent, a revisionist.⁶ All philosophical theories depart to some degree, in some way, from the particulars of folk thinking. In reply: we

need to be careful to track the distinctive nature of revisionist theories. What makes a theory revisionist in the sense adumbrated here is not merely the facture of a departure from commonsense. Instead, it is the fact that the truth of the proposal is in *conflict* with commonsense. To be sure, conflict can come in degrees, so theories can be more and less revisionist. We have also seen that the target of revision can vary. For example, as noted above, one might be revisionist about our understanding of the nature of free will without being a revisionist about particular practices that implicate free will.

Whether any revisions are interesting or philosophically salient partly depends on the background presumptions about how common conflicts with commonsense are expected to be in philosophical theorizing. Within normative ethics and philosophy of physics, such conflicts are a matter of course. So, revisionism is less notable. Within philosophy of mind, the gradual retreat from eliminativism to revisionism about folk psychological terms took longer, despite the widespread influence of Quinean holism. However, within contemporary debates about free will – that is, a literature still heavily structured by the methods and presumptions of 1960s and 1970s analytic metaphysics – revisionism remains a minority position.

The distinctiveness of revisionism is more apparent if one considers the history of the subfield. For example, it would be grossly uncharitable to interpret most of the historical proponents of the conditional analysis as aspiring to reform meaning. Similarly, to assert that most current compatibilist accounts are revisionist in the sense under consideration would also be to say that most contemporary compatibilist accounts are entirely unresponsive to the burdens of motivating their revisions, of explaining why they aren't changes in topic, and in acknowledging that folk thinking is indeed incompatibilist in significant measure. These would be serious lacunae in contemporary compatibilist accounts. The fact of their absence in virtually all compatibilist accounts that do not explicitly identify as revisionist would seem to justify incompatibilist complaints that compatibilism is a wretched subterfuge.

It is more plausible to hold that revisionist compatibilists really are up to something different than traditional compatibilists.

5 Recent Issues in Revisionist Theorizing

Two recent debates about the forms revisionist theories can take – a matter of the conceptual space available to revisionists – merit some discussion. One is Nichols's argument for the possibility of a view he calls *discretionism* and a second is McCormick's argument for restricting revisionism to denotational revisionism, or a kind of *replacementism*. Nichols's approach might be thought of as an attempt to expand the scope of options available to revisionists whereas McCormick has argued for a narrower range of options available to the revisionist.

Nichols's version of revisionism about free will and moral responsibility is subtle and innovative. As we saw above, he argues that folk thinking has a robust and readily identifiable commitment to incompatibilist conceptions of agency. However, he maintains that for moral and pragmatic reasons, the attitudes and practices we associate with responsibility (including moral anger and retributive impulses) should be retained. What makes the account revisionist – as opposed to fictionalist or illusionist (as in Smilansky 2000) – is that he thinks that our incompatibilist intuitions can be severed from those practices, so that our practices retain adequate normative standing. His is a kind of conceptual revisionism with little or no revisionism about the practice.

One innovation of Nichols's account is that he endorses a kind of discretionism, a view according to which variability in the reference conventions of the target term permit one to selectively adopt eliminativist discourse in one moment, and non-eliminativist discourse in

another. The underlying idea is that some terms exhibit distinct reference conventions (e.g., having both a causal-historical reference convention *and* a distinct descriptivist convention). Nichols thinks “free will” and “moral responsibility” are such terms (2015, p. 166). What makes the account discretionist is that one is licensed or permitted to select among these conventions at will – at least in some contexts.⁷ This referential flexibility is appealing, Nichols thinks, because it affords us particular psychological advantages. During one’s long dark nights of the soul, discretionism allows us to remind ourselves that we are not free agents of the sort that figure in libertarian accounts of free will, and, thus, that we cannot be responsible for our failings in the way characterized by those accounts. In daily life, however, it allows us to rightly insist that people can be blameworthy and that they can deserve our condemnation.

Whether discretionism is a stable view remains a matter of some dispute (see McCormick 2017; Vargas 2017; for a reply see Nichols 2017). Here is one concern: even if there are multiple (non-privileged) reference conventions for “free will” and “moral responsibility,” it isn’t clear that referential discretionism gets us anything different than a non-discretionist form of revisionism. After all, the non-discretionist revisionist can already say (in chorus with the discretionist) that one is not free in the way specified by the incompatibilist – i.e., that one does not have the properties that the incompatibilist says are required for free will and moral responsibility. So, the discretionist is not thinking a new thought on those long, dark nights of the soul when she concludes that she is not free and responsible as specified by a given incompatibilist account. Moreover, the sense in which the discretionist is putatively joining the eliminativist in thinking that one is not morally responsible cannot be the same sense in which an eliminativism worth the name makes that claim. On pain of there being insufficient reason to eliminate the practice and discourse of freedom and responsibility (and, thus, to be a distinctive view), eliminativism must deny that our current freedom and responsibility-characteristic attitudes and practices are in good normative standing. Nichols’s discretionist, however, does not (and cannot) think *that*. The entire point of Nichols’s positive account just is a defense of integrity and value of those attitudes and practices. So, the virtues of discretionism as a distinctive form of revisionism, or perhaps as an alternative to familiar forms of revisionism, remain unclear.

In a series of recent papers, Kelly McCormick (2013, 2015, 2016) has deepened our understanding of the resources and challenges facing revisionist accounts of free will and moral responsibility, and developed new resources for responding to a range of concerns about revisionist approaches to theorizing. If she is right, connotational revisionism is unstable. On her account, the only viable form of revisionism must involve the relatively ambitious project of re-anchoring the referent of “free will” and “moral responsibility” in some nearby property.

(To recap: the difference between denotational and connotational revisionism is the difference between revisionist accounts according to which error-ridden folk commitments can be jettisoned without affecting reference (connotational revisionism) and revisionist accounts according to which error-ridden folk commitments *cannot* be jettisoned without some kind of reference change (denotational revisionism). Put differently, denotational revisionism concedes that the target term has no extension, but goes on to recommend a reference shift. Connotational revisionism holds that the target term has an extension, but that we have had false beliefs about the extension.)

McCormick argues that connotational revisionism faces a kind of dilemma: either the to-be-revised-away content is widespread (and, thus, it is implausible to insist that it is neither constitutive nor essential), or it is insufficiently widespread to ground a substantive conflict between the diagnostic and prescriptive elements of the account (2016, p. 117). For this reason, she favors a denotational analysis. However, denotational revisionism faces challenges of its own. First, one might doubt that denotational revisionism is distinct from eliminativism. Second, there is some reason to be skeptical whether connotational revisionism will go quietly into the night.

Is denotational revisionism really eliminativism in all but name? Notice that the denotational revisionist must grant that the pre-revised term lacks an extension. That fact, one might think, suffices for showing that any denotational revisionism is eliminativist. If revisionism just is restricted to denotational revisionism, then revisionism collapses into eliminativism, and there is nothing left that could count as a revisionist theory of free will.

McCormick allows that there is a sense in which denotational revisionism is indeed a kind of eliminativism. This result is misleading, however, because there remains a substantive difference between the position of the eliminativist and the denotational revisionist: the denotational revisionist is ultimately a kind of success-theorist, holding that there is reason to re-anchor talk in a nearby property (for example, one that does not involve powers invoked by libertarian accounts). Unlike the eliminativist, when we implement the revisionist's proposal, we *will* successfully refer. Moreover, our referring to this new property will preserve the characteristic structure of practices, attitudes, and judgments at stake in our concern for whether we are free (or responsible). So, a successful denotational revisionism might just as well be labeled *replacementism*, she notes, thereby highlighting that the term and its linguistic and social functions remain intact in light of identifying a suitable alternative property to anchor them.

Here's one way to see the point: there is a big difference between patching a boat with new parts, and dis-assembling the boat because one no longer regards it as seaworthy. As long as the characteristic function of the term is preserved, patching the function with a different referent remains a reasonable strategy. Thus, so long as we retain the plausible social or psychological functions of terms like "free will" and "moral responsibility," there will be a substantive difference between revisionist and eliminativist proposals.

As a matter of the sociology of those who self-describe as eliminativists and revisionists, McCormick's observation about the aspirational differences between eliminativism and revisionism is surely right. Yet, a puzzle lurks. Consider the point made at the outset of this chapter, about how one might be a revisionist about true love, yet ultimately conclude that true love does not exist.⁸ A similar situation seems possible for free will. A theorist might appeal to an adequate replacement property to ground a proposed reference shift. This would suffice to make the theorist a denotational revisionist. However, that theorist might later conclude that there is a further reason for thinking that the new target property also fails to obtain. If no further revision seems viable, eliminativism is the order of the day. Thus, it seems, the difference between revisionism and eliminativism, if there is one, is not the difference between a success theory and an error theory.

Suppose McCormick grants that eliminativism and revisionism can, in fact, overlap in the particular way exemplified in the true love case. She would still be right that the ambitions of the proposals remain distinct, even in the rare limit case where a theory is both eliminativist and revisionist. Even in the true love case, the revisionist undertakes her revision with the thought that it is promising as a success theory; this is not true of the eliminativist. Even if eliminativism ends up being the right account, this is a further discovery. Despite the fact that revisionist theories are picked out by the conflict between folk and all-things-considered theorizing – and although they are ordinarily undertaken with an aspiration of providing a success theory, and the best accounts will give some reason for thinking things are as the account prescribes – whether reality cooperates remains a further thing.

A second family of challenges for McCormick's account concerns whether the argument against connotational revisionism succeeds. Here is one reason for doubt. Suppose "free will" is a natural kind (Heller 1996; Deery 2021) – or at least a *genuine kind* (see Spencer 2012). On such an account, free will refers to property not fixed by a description but instead by a suitable

causal-historical chain. If that is the sort of term “free will” is, then discourse about free will can tolerate considerable and widespread errors about the nature of free will and still refer. Similarly, talk and beliefs about water can refer despite beliefs that it is, for example, one of the four basic elements of the world. To be sure, this line of response depends on moral responsibility being a particular kind of thing. At the same time, it does seem a kind of cost to McCormick’s argument against connotational revisionism that it requires that free will is not a natural or genuine kind.

Is connotational revisionism possible without free will being a genuine kind, or being picked out by a causal-historical chain? There are a variety of descriptivist strategies for accommodating false beliefs in such a way that reference is retained even in the face of some amount of false beliefs about the nature of the thing in question (Cf. Lewis 1983). On these accounts, so long as enough of the description of free will (or moral responsibility, or ...) is true, or there is some property that renders enough of the platitudes true, jettisoning some false beliefs need not disrupt reference. If reference can be preserved, even in the face of the falsity of some platitudes under descriptivist reference conventions, then there is little reason to think widespread commitments cannot also be jettisoned without loss of the referent. In sum, the possibility of connotational revisionism does seem not turn on whether free will is a genuine kind, referred to by a causal-historical chain (Cf. Vargas 2013a, pp. 91–96).

6 Further Developments

Perhaps the most visible unresolved issues facing revisionist theories concern the role of revisionism about free will in connection with some normative ideas. This concluding section briefly canvasses three issues where revisionist work has interesting implications for recent debates. They include the relationship of free will to moral responsibility, disputes about desert, and the basis on which one might decide between revisionism and eliminativism.

As noted at the outset, a fair number of self-described revisionist accounts seem to be focused on issues that are readily construed as centrally about moral responsibility, rather than free will. Given the varied conceptions of what kind of thing free will is, how it is to be characterized, and the diverse roles played by the term in popular and technical discourse (Vargas 2011, pp. 152–155, 2013b, pp. 179–181), one can imagine that some theorists will be tempted to be eliminativist about free will and revisionist about responsibility.

Within the literature on moral responsibility, desert has become an increasingly important locus of disagreement. This was an issue noted by Smart (1961), Wallace (1994, p. 228), and Scanlon (1998, p. 275), but only recently has it become a matter of sustained discussion (see McKenna 2009, 2012, pp. 114–120; Roskies and Bertram Malle 2013; Nahmias 2014, p. 55 n.3; Doris 2015b; Caruso and Morris 2017). In particular, for theories that seek to ground the normative basis of responsibility in broadly consequentialist justifications (theories that in recent years have tended to be revisionist), it is an open question whether such accounts have the resources to capture our ostensibly commonsense concern for desert, and, if not, whether desert itself admits of a revisionist reconstruction (see McGeer 2012, 2015; Vargas 2015, as well as McCormick in this volume).

The fate of revisionist theories is partly tied to one’s conception of the aspirations for philosophical theories of free will and moral responsibility. Roughly speaking, there seem to be two competing impulses: (1) the desire to rescue our default conceptual or theoretical commitments and (2) the aspiration to ground the integrity of practices, attitudes, or other relatively concrete phenomena. For many eliminativists, the impossibility of vindicating ordinary theoretical commitments is the basis for the eliminativism and the basis for rejecting the

possibility of grounding our practices. In contrast, for revisionists, the significance of rescuing our naïve theoretical commitments pales in comparison to the significance of grounding our practices. By such lights, the eliminativist's concern for vindicating folk conceptions of agency, free will, desert, and moral responsibility is like a hungry person's fascination with accurate portraits of food. It is an explicable interest, but the wrong preoccupation when there is real food readily to eat.

Even so, eliminativism may yet retain some allure. Suppose that we conclude that primary philosophical stakes are practical and not theoretical – that is, they concern some subset of identifiable judgments, practices, and attitudes, and *not* the contents of our naive theories about the foundations of these things. Even if we decide that revisionists can provide an adequate ground for those practices, we might still ask whether, all things considered, we do well to keep these things at all. Might we be better off without the responsibility-characteristic practices, attitudes, and judgments? Eliminativism might yet be the better alternative.

One reply is the one given by P.F. Strawson (1962): it involves holding that the underlying attitudes implicated in responsibility practices are simply implastic pieces of our psychological architecture. Many eliminativists, and at least some revisionists, are inclined to regard this sort of claim with suspicion. There is a live issue here about how to measure the benefits of free will and responsibility practices, attitudes, and judgments, and what the metric should be for deciding whether we do better to keep or eliminate them.

Eliminativists (like Pereboom and Caruso) will extol the putative benefits of a world without free will and the attendant responsibility practices. Revisionists who think there are important normative and social payoffs of retaining these things (e.g., McGeer, Nichols, Vargas) will insist that the benefits of systematically foregoing responsibility have been overstated, and indeed, that the purported benefits of eliminativism are at best local, and often byproducts of free-riding on social practices structured by moral responsibility. Gandhi's and Martin Luther King's apparent lack of resentment, indignation, and retributivism is undoubtedly appealing, and this fact may seem to favor the psychological picture advocated by eliminativists. To be sure, pacificism and diminished reactivity may seem particularly appealing when the alternative is violent, angry, and vindictive. Yet a revisionist reply remains available: the appeal and effectiveness of non-blaming exemplars may depend in part on the lurking alternative of blame-propelled practices. Absent the alternative of blame-based practices, and especially at the level of stable, widespread human practices, it is unclear whether our admiration for non-blaming stances could be sustained. If so, then eliminativism's moral appeal is cashing a check that draws on blame's account.⁹

Notes

- 1 Dennett (1984, 2003) has an important but ambiguous relationship to revisionism. Although his account is associated with the slogan “the varieties of free will worth wanting” (the subtitle of his first book on free will), and perhaps popularly thought of as an early proponent of revisionism, the actual arguments and the views advanced in his accounts tend to overlook the differences in revisionism and conventional compatibilism. Moreover, when pressed, Dennett has tended to disavow the revisionist reading of his work. For a more detailed discussion of the difficulty of treating Dennett as a conventional revisionist, see Vargas (2005b).
- 2 The nature of free will and its characterization is a contentious issue. Some are inclined to provide a core characterization in terms of moral responsibility (as in Fischer et al. 2007), some in terms of desert (Pereboom 2001), some in terms of deserving retribution (Caruso and Morris 2017). Whether these are mistakes is not something on which this discussion needs to take a stand. For some relevant discussions, though, see van Inwagen (2008), Balaguer (2010), Vargas (2011), and Franklin (2015).

- 3 Compare McCormick's formulation of revisionism: "Revisionism is the view that we can and should distinguish between what we think about moral responsibility and what we ought to think about it, that the former is in some important sense implausible and conflicts with the latter, and so we should revise our concept accordingly" (2016, p. 109).
- 4 Normally, denotational revision is characterized by the idea that there is some nearby property that is a suitable basis for various practices, judgments, and attitudes that are impugned by one's side diagnostic account, and the revisionist proposal is a recommendation to shift reference to explicitly track that nearby property. So, this would be a special case of reference shift, from an indeterminate referent to a determinate referent.
- 5 Doris disassociates free will from moral responsibility (2015a, pp. 9–11), and says little about free will specifically. So, his account does not obviously entail revisionism about free will. Given what he says about reflectivism and agency, however, it is not difficult to imagine recasting most of his arguments against folk conceptions of responsibility as arguments against folk conceptions of free will.
- 6 This objection was rightly raised by Michael McKenna (2009) against previous formulations of revisionism (e.g., Vargas 2005a) where revisionism was wrongly characterized in terms of any departure from commonsense commitments.
- 7 Were there a uniformly privileged reference convention, it is less clear that discretionism is permissible in the same way.
- 8 Here, as there, we can ignore a complexity about the difference between a shared concept, term, or practice and whatever the individual correlates of these things may be. Revisionism of the sort discussed here concerns proposals about the former (collective) versions of these things, so any example about an individual case will be, to that extent, modestly misleading.
- 9 A limitation of this chapter is that it was submitted in 2016. As such, it does not engage with the subsequent literature that grew up around conceptual ethics or conceptual engineering. My thanks to the editors Gregg Caruso, Kelly McCormick, and Stephanie Vargas for feedback on prior drafts of this essay.

Bibliography

- Arneson, R.J. (2003). The smart theory of moral responsibility and desert. In: *Desert and Justice* (ed. S. Olsaretti), 233–258. Oxford: Oxford University Press.
- Balaguer, M. (2010). *Free Will as an Open Scientific Problem*. Cambridge, MA: MIT Press.
- Bennett, J. (1980). Accountability. In: *Philosophical Subjects* (ed. Z. Van Straaten), 14–47. New York: Clarendon Press.
- Björnsson, G. and Persson, K. (2012). The explanatory component of moral responsibility. *Nous* 46 (2): 326–354.
- Björnsson, G. and Persson, K. (2013). A unified empirical account of responsibility judgments. *Philosophy and Phenomenological Research* 87 (3): 611–639.
- Bourget, D. and Chalmers, D.J. (2014). What do philosophers believe? *Philosophical Studies* 170 (3): 465–500.
- Caruso, G. (2015). Free will eliminativism: reference, error, and phenomenology. *Philosophical Studies* 172 (10): 2823–2833.
- Caruso, G.D. and Morris, S.G. (2017). Compatibilism and retributive desert moral responsibility: on what is of central philosophical and practical importance. *Erkenntnis* 82 (4): 837–855.
- Collins, R. (1998). *Sociology of Philosophies*. Cambridge, MA: Belknap.
- Deery, O. (2015). The fall from Eden: why libertarianism isn't justified by experience. *Australasian Journal of Philosophy* 93 (2): 319–334.
- Deery, O. (2021). Free actions as a natural kind. *Synthese* 198 (1): 823–843.
- Dennett, D. (1984). *Elbow Room*. Cambridge: MIT.
- Dennett, D. (2003). *Freedom Evolves*. New York: Viking.
- Doris, J. (2002). *Lack of Character*. New York: Cambridge University Press.

- Doris, J. (2015a). *Talking to Our Selves: Reflection, Ignorance, and Agency*. New York: Oxford University Press.
- Doris, J. (2015b). Doing without (arguing about) desert. *Philosophical Studies* 172 (10): 2625–2634.
- Doris, J., Knobe, J., and Woolfolk, R.L. (2007). Variantism about moral responsibility. *Philosophical Perspectives* 21 (1): 183–121.
- Double, R. (1996). *Metaphilosophy and Free Will*. New York: Oxford University Press.
- Faucher, L. (2016). Revisionism and moral responsibility for implicit attitudes. In: *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics* (ed. M. Brownstein and J. Saul), 115–145. Oxford: Oxford University Press.
- Fischer, J.M., Kane, R., Pereboom, D., and Vargas, M. (2007). *Four Views on Free Will*. Malden, MA: Wiley-Blackwell.
- Franklin, C. (2015). Everyone thinks that an ability to do otherwise is necessary for free will and moral responsibility. *Philosophical Studies* 172 (8): 2091–2107.
- Heller, M. (1996). The mad scientist meets the robot cats: compatibilism, kinds, and counterexamples. *Philosophy and Phenomenological Research* 56: 333–337.
- Hurley, S. (2000). Is responsibility essentially impossible? *Philosophical Studies* 99: 229–268.
- Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. New York: Oxford University Press.
- Knobe, J. and Doris, J. (2010). Responsibility. In *The Moral Psychology Handbook* (ed. J. Doris and The Moral Psychology Research Group), 321–354. Oxford: Oxford University Press.
- Levy, N. (2011). *Hard Luck: How Luck Undermines Freedom*. New York: Oxford University Press.
- Lewis, D. (1973). *Counterfactuals*. Cambridge: Harvard University Press.
- Lewis, D. (1983). How to define theoretical terms. In: *Philosophical Papers*, 78–95. New York: Oxford University Press.
- McCormick, K. (2013). Anchoring a revisionist account of moral responsibility. *Journal of Ethics and Social Philosophy* 7 (3): 1–19.
- McCormick, K. (2015). Companions in innocence: defending a new methodological assumption about moral responsibility. *Philosophical Studies* 172 (2): 515–533.
- McCormick, K. (2016). Revisionism. In: *Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), 109–120. New York: Routledge.
- McCormick, K. (2017). Why we should(n't) be discretionists about free will. *Philosophical Studies* 174 (10): 2489–2498.
- McGeer, V. (2012). Co-reactive attitudes and the making of moral community. In: *Emotions, Imagination, and Moral Reasoning* (ed. R. Langdon and C. Mackenzie), 299–326. New York: Psychology Press.
- McGeer, V. (2015). Building a better theory of responsibility. *Philosophical Studies* 172 (10): 2635–2649.
- McKenna, M. (2009). Compatibilism and desert: critical comments on four views on free will. *Philosophical Studies* 144: 3–13.
- McKenna, M. (2012). *Conversation and Responsibility*. New York: Oxford University Press.
- Nahmias, E. (2006). Folk fears about freedom and responsibility: determinism and reductionism. *Journal of Culture and Cognition* 6 (1–2): 215–238.
- Nahmias, E. (2011). Intuitions about free will, determinism, and bypassing. In: *Oxford Handbook of Free Will*, 2e (ed. R. Kane), 55–576. New York: Oxford University Press.
- Nahmias, E. (2014). Response to Misirlisoy and Haggard and to Björnsson and Pereboom. In: *Moral Psychology Volume 4: Free Will and Moral Responsibility* (ed. W. Sinnott-Armstrong), 43–57. Cambridge, MA: MIT Press.
- Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research* 73 (1): 28–53.
- Nahmias, E.D., Coates, J., and Kvaran, T. (2007). Free will, moral responsibility, and mechanism: experiments on folk intuitions. *Midwest Studies in Philosophy* XXXI: 214–241.
- Nelkin, D.K. (2011). *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.
- Nichols, S. (2006). Folk intuitions on free will. *Journal of Cognition and Culture* 6 (1 & 2): 57–86.
- Nichols, S. (2007a). After incompatibilism: a naturalistic defense of the reactive attitudes. *Philosophical Perspectives* 21: 405–428.

- Nichols, S. (2007b). The rise of compatibilism: a case study in quantitative history of philosophy. *Midwest Studies in Philosophy* 31: 260–270.
- Nichols, S. (2015). *Bound: Essays on Free Will and Responsibility*. New York: Oxford University Press.
- Nichols, S. (2017). Replies to Kane, McCormick, and Vargas. *Philosophical Studies* 174 (10): 2511–2523.
- Nichols, S. and Knobe, J. (2007). Moral responsibility and determinism: the cognitive science of folk intuitions. *Nous* 41 (4): 663–685.
- Nozick, R. (1981). *Philosophical Explanations*. Oxford: Clarendon Press.
- Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.
- Roskies, A. and Bertram Malle, F. (2013). A Strawsonian look at desert. *Philosophical Explorations* 16 (2): 133–152.
- Roskies, A. and Nichols, S. (2008). Bringing responsibility down to earth. *Journal of Philosophy* 105 (7): 371–388.
- Scanlon, T.M. (1998). *What We Owe to Each Other*. Cambridge, MA: The Belknap Press of Harvard University Press.
- Singer, I. (2002). Freedom and revision. *Southwest Philosophy Review* 18 (2): 25–44.
- Smart, J.J.C. (1961). Free will, praise, and blame. *Mind* 70: 291–306.
- Smilansky, S. (2000). *Free Will and Illusion*. New York: Clarendon Press.
- Spencer, Q. (2012). What ‘biological racial realism’ should mean. *Philosophical Studies* 159 (2): 181–204.
- Strawson, G. (1986). *Freedom and Belief*. Oxford: Oxford University Press.
- Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies* 75: 5–24.
- Strawson, P.F. (1959). *Individuals, an Essay in Descriptive Metaphysics*. London: Methuen.
- Strawson, P.F. (1962). Freedom and resentment. *Proceedings of the British Academy*. XLVIII: 1–25.
- van Inwagen, P. (2008). How to think about the problem of free will. *Journal of Ethics* 12: 327–341.
- Vargas, M. (2004a). Responsibility and the aims of theory: Strawson and revisionism. *Pacific Philosophical Quarterly* 85 (2): 218–241.
- Vargas, M. (2004b). Libertarianism and skepticism about free will: some arguments against both. *Philosophical Topics* 32 (1&2): 403–426.
- Vargas, M. (2005a). The revisionist’s guide to responsibility. *Philosophical Studies* 125 (3): 399–429.
- Vargas, M. (2005b). Compatibilism evolves? on some varieties of Dennett worth wanting. *Metaphilosophy* 36 (4): 460–475.
- Vargas, M. (2011). The revisionist turn: reflection on the recent history of work on free will. In: *New Waves in the Philosophy of Action* (ed. J. Aguilar, A. Buckareff and K. Frankish), 143–172. New York: Palgrave Macmillan.
- Vargas, M. (2013a). *Building Better Beings: A Theory of Moral Responsibility*. Oxford, UK: Oxford University Press.
- Vargas, M. (2013b). If free will does not exist, then neither does water. In: *Exploring the Illusion of Free Will and Moral Responsibility* (ed. G. Caruso), 177–202. Lanham, MD: Lexington Books.
- Vargas, M. (2013c). Situationism and moral responsibility: free will in fragments. In: *Decomposing the Will* (ed. T. Vierkant, J. Kiverstein and A. Clark), 325–349. New York: Oxford University Press.
- Vargas, M. (2015). Desert, responsibility, and justification: reply to Doris, McGeer, and Robinson. *Philosophical Studies* 172 (10): 2659–2678.
- Vargas, M. (2016). The Runeberg problem: theism, libertarianism, and motivated reasoning. In: *Libertarianism and Free Will: The Interplay of Religious Belief and Free Will* (ed. K. Timpe and D. Speak), 27–47. New York: Oxford University Press.
- Vargas, M. (2017). Implicit bias, moral responsibility, and moral ecology. In: *Oxford Studies in Agency and Responsibility* 4 (ed. D. Shoemaker), 219–247. New York: Oxford University Press.
- Wallace, R.J. (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
- Walter, H. (2004). Neurophilosophy of moral responsibility: the case for revisionist compatibilism. *Philosophical Topics* 32 (1&2): 477–532.
- Weigel, C. (2013). Experimental evidence for free will revisionism. *Philosophical Explorations* 16 (1): 31–43.

Part III

The Science of Free Will

How the Laws Constrain Causation, Counterfactuals, and Free Will

KADRI VIHVELIN

UNIVERSITY OF SOUTHERN CALIFORNIA

The free will/determinism problem is usually presented as a problem about the compatibility of laws of a certain kind – deterministic – with a special property of human beings: free will. And many people think, sometimes on the basis of argument, sometimes because it just seems obvious to them, that determinism is incompatible with free will. Some say that determinism rules out not just free will but also moral responsibility. Others disagree, claiming that we might be morally responsible even if deterministic laws rule out all alternative possibilities and thus all free will. And so the debate goes, always for very high stakes because free will and moral responsibility matter a great deal to us.

I want to take a different approach. I think the problem is more general. I think that the free will/determinism problem is, most fundamentally, a problem about the (apparent) conflict between the way things are at a deterministic world and our commonsense beliefs about causation, our causal powers, and the causal powers of ordinary objects like hammers and baseballs. To make progress on the free will/determinism problem, we must understand causation better. To understand causation better, we must understand the relation between causation, counterfactuals, and the laws.

I think we should be surprised by how quick we are to conclude that the truth of determinism means the absence of free will. After all, we already know that we are physical beings in a physical world, subject to the same laws that apply to everything else in the universe. We know that the laws set limits on our abilities and on our powers of choice and action. We accept these limits while still believing that we are agents with free will. Yet determinism, when explained in a philosophy or science classroom, tends to make us think we have no free will.

We need an explanation for why we move so quickly from our cheerful acceptance of the fact that our powers of free will and agency are limited without being eliminated by the physical world and its laws to the conclusion that *if these laws turn out to be deterministic then we have no free will at all*.

I'm a compatibilist.¹ I think that we do have free will – and by this I mean, among other things, that we have causal powers of choice and action and that we are often in situations where *we are able to choose and act otherwise*. I think that there are no good arguments for the dire conclusion that the truth of determinism means the absence of free will and the absence of these causal powers. But I do not deny that incompatibilist intuitions are powerful.

In his classic defense of compatibilism, R.E. Hobart famously said that our incompatibilist intuitions are due to “our natural want of the analytic imagination.”² He compared our failure to understand how we might be free-willed agents at a deterministic world to the failure of a group of village folk to understand how a steam engine works. Even after they had listened to a long lecture on how the parts fit together to make the engine work, they asked if there was a horse inside.

There is clearly something to Hobart's diagnosis. But I don't think that's all there is to it. Determinism is a thesis about the laws. It's not a thesis about causation or causal explanation. And yet a very common response to determinism is to think that it undermines our commonsense beliefs not only about free will but also about causation. Why is this? I will be arguing that the answer lies in an assumption we make about counterfactuals.

That's the first part of this paper. The second part is an argument that we should reject the current dominant theory of counterfactuals – David Lewis's theory³ – in favor of a theory that holds the laws fixed. I will be recommending the Simple theory once defended by Jonathan Bennett.⁴ I don't claim that the Simple theory makes it easier to defend compatibilism. My claim is that we should prefer the Simple theory because it's a better theory of counterfactuals. But I will also argue that the Simple theory provides us with the tools we need to think more clearly about the free will/determinism problem.

I will begin by saying something about how we acquire our beliefs about our causal powers.

1 How Experience Teaches

When we try to decide what to do, we assume that we are choosing from some range of alternatives. We believe that we are able to choose any of these alternatives and we believe that our choice will, *ceteris paribus*, be causally efficacious.

Even when we aren't trying to make a decision, we think of ourselves as agents with the power to do things that make a difference to the world around us. Our powers are limited by our skills and physical capacities and by the situations in which we find ourselves. And, of course, by the laws of nature and the past. We can't make bread from stone, part the Red Sea, or run faster than the speed of light. And we can't do anything to change the past, including those actions we now regret. But these limits leave us free to move our bodies in ways that cause some things to happen and prevent other things from happening. By throwing a baseball, you are able to break a window. By catching the ball before it reaches the window, I am able to prevent the window from breaking. We think of these powers of causing and preventing as being under our control in the following way: When we cause or prevent an outcome, we didn't have to do what we did; what we did was *up to us*; we were *able to do otherwise*; doing otherwise was *in our power*; we *had a choice*.

We also believe that we sometimes have control over outcomes that we ourselves do not cause. To continue with the baseball story, suppose that I see you lift the baseball and take aim at the window. I am standing between you and the window, catcher's mitt in hand, and I have plenty of time to raise my mitt and intercept the ball. But I think it would be fun to watch the window breaking, so I don't. I don't cause you to throw the ball, and it doesn't seem right to

say that I cause the window to break. But it does seem right to say that I *allowed* it to break. And part of what we believe, when we believe this, is that I had the *power to do something that would have prevented the window from breaking*.

How do we acquire these beliefs about our powers? Here's my small, not very controversial, conjecture: We acquire these beliefs at the same time, and in the same way, that we acquire our beliefs about causation more generally.⁵ A baby uses her hands to pick up a rattle and discovers that when she shakes the rattle, it makes a certain kind of noise. She moves her hands a certain way, the noise starts; she stops moving them that way, the noise stops. She shakes her teddy bear; no noise. She squeezes the teddy bear; a different kind of noise. And the baby thereby learns both that the teddy bear and the rattle make these kinds of sounds when manipulated a certain way *and* that she is able to produce these sounds by manipulating them in these ways. By experimenting in these ways, we gradually learn the truth of counterfactuals of the following kind: "If I had not moved my body like *that*, that object (that rattle, that teddy bear) would not have done *that*."

As we learn more about the causal powers of the particular objects around us and, at the same time, about our own causal powers, we acquire beliefs in more counterfactuals, and we progress from beliefs about particular objects to counterfactual generalizations about similar objects. We acquire this counterfactual knowledge not only by noticing which manipulations produce changes in which kinds of objects, but also by noticing how objects of certain kinds persist unchanged, or change on their own, in the absence of our interventions. We learn that scissors are good at cutting paper, not so good at cutting cloth or wood. We learn that balls tend to roll down sloped surfaces unless blocked by some other object. And so on.⁶

As our ability to predict the characteristic behaviors of objects increases, we acquire the ability to prevent things from happening. We see a ball rolling down a hill; we put our foot in front of it, preventing it from reaching the bottom. We see a plate teetering on a counter, and grab it before it falls and breaks. An older child might realize that a younger child is in danger, and reach out and grab the child before she falls or touches something sharp or hot. And once we understand what it is to prevent something from happening – once we understand how to anticipate an outcome that would have happened, had we not intervened – we can also understand what it is to *allow* an outcome. For to allow an outcome is to fail to prevent some outcome that we *could have prevented*. More specifically, we allow an outcome if we were able to move our bodies in some way such that, *if we had moved them that way, then and there, that outcome would not have happened*.

These are some of our commonsense beliefs about the ways in which we are actual and potential causes of a wide variety of events and states of affairs.⁷ The most interesting metaphysical commitment of these beliefs is that we have, *somehow or other*, causal powers that persist during times that we are not exercising them. This belief is a highly robust one, and one that is embedded in our ordinary talk of allowing various kinds of things to happen – undesirable consequences, our own actions, the evolution of our future selves. But it's also undeveloped, in crucial respects, in the way that commonsense beliefs tend to be. Nothing that I've said supports the claim that commonsense has a metaphysical view about the nature of these powers or about what the laws must be like in order for us to have these powers. Nothing that I've said supports the claim, famously made by Roderick Chisholm,⁸ that our agential powers are like the power "which some would attribute only to God: each of us, when we act, is a prime mover unmoved." So far as commonsense is concerned, it might be that our causal powers require determinism or indeterminism or are compatible with both.⁹

2 Determinism and the Three Stages of Denial

And yet when we learn what the thesis of determinism says, and are asked to take seriously the thought that science might tell us that determinism is true, we respond with horror and what I call the *three stages of denial*.

Determinism is defined in the literature in one of two ways. The Entailment definition says that determinism is the thesis that for every time t , there is some proposition S about the state of the world at t which, together with some proposition L about the laws of nature, entails every true proposition about the state of the world at every other time.¹⁰ The Modal definition says that determinism is the thesis that, for any two worlds w_1 and w_2 that satisfy the same set of laws L as our world, and for every time t , if w_1 and w_2 are alike at t , they are alike at all other times; if w_1 and w_2 are different at t , they are different at all other times.¹¹ The Modal definition is stronger; if a world is deterministic according to the Modal definition, then it is also deterministic according to the Entailment definition, but not vice versa.¹² But it doesn't matter which definition is used, so long as it is clearly explained.¹³

The first stage of denial is the Incredulous Stare: Determinism *can't* be true; it literally makes no sense. The second stage is Misinterpretation and Epiphenomenalism: Determinism is understood as *almost-determinism*: all events are governed by deterministic laws *except* for our choices. If almost-determinism were true, we would still have free will but no freedom of action. We would still be able to deliberate and make choices, but none of our choices would make any causal difference, not even to the movement of our own bodies. The third stage is Fatalism: Everything that happens – including our own thoughts, choices, and efforts – is necessary; everything that doesn't happen is impossible. We have no free will and no causal power. It isn't really *us* – or for that matter, horses or steam engines or baseballs – that does the causing. It's "just the laws making everything happen."

I'm not making this up. I've been teaching the free will/determinism problem in an introductory semester-long class to undergraduates for almost twenty years. It took me awhile to recognize it, but I am now confident that I am correctly reporting a common reaction.

The shock and horror come in the move from a claim I call the No Exception thesis to either the Entailment or the Modal definitions of determinism. The No Exception thesis says that we are *no exception* to the laws that govern everything else in the universe: the laws apply to us in the same way that they apply to everything else in the universe. The No Exception thesis does not provoke any of the three stages of denial. It is the explanation of the difference between a deterministic and non-deterministic system of laws, and, especially, the difference between a deterministic and non-deterministic world, that triggers the panic.

Why this response?

Some readers might not be persuaded that there is anything problematic because they already think that some more sophisticated version of the third stage is the correct response. So I will say a bit about why I think this conclusion is not warranted.

3 How the Laws Constrain

We begin our philosophical theorizing with some commonsense data. Here is one datum about the laws of nature, which any adequate theory of laws must accommodate, somehow or other.

The laws constrain us – and everything else in the universe – in a substantive way. There are some things we are not able to do *because of the laws*. Knowing what the laws are gives us

useful, non-trivial information about what we can do. This is distinct from two other kinds of constraints – the constraints of logic, broadly understood, and the constraints that distinguish individuals from one another.

It is logically impossible to: square a circle, stay home all day while also leaving the house that day, bake a cake bigger than itself. We are not able to do or cause any of these things. If we tried, we would fail, no matter how hard we tried, no matter how many times we tried, no matter how ingenious our efforts were. Nor would we be able to do something logically impossible if the laws of nature were different; our inability is due to logic (broadly understood), not the laws.

It isn't *logically* impossible for me to speak Chinese, walk a tight-rope without falling more or less immediately, or run a mile in less than five minutes. Nor do the laws of nature rule this out, for the world is filled with Chinese speakers, tight-rope walkers, and fast runners. And I might even acquire these abilities myself, some day. But I'm nevertheless not able to do these things *now*. If I tried, I would fail, and my failure wouldn't be a mere fluke or something we might write off as bad luck. I would fail *no matter how hard I try, or how many times I try*. I would fail because I don't, *given the way I am now*, have what it takes to do any of these things. Given the way I am now, the laws preclude my doing any of these things.

It isn't logically impossible for me to accelerate to a speed faster than the speed of light, but the laws say that no one – and nothing – does. And because of this no human being is able to win a race with a light-beam, *no matter what*. It makes no difference how good a runner you are, how long you train, or how hard you try. Your failure is actually and *counterfactually* guaranteed, by the laws. There are possible worlds where the laws of physics are different, and even different enough to permit faster than light acceleration. And there may be possible worlds where you are able to run faster than the speed of light. But the laws that govern the actual universe dictate that you will never acquire this ability. Nor are you able to acquire this ability. There is nothing that you are now able to do – in the way that you are now able to take a Chinese course, join a circus class, or start running every day – such that, if you did it long enough, and seriously enough, you would eventually acquire the ability to run faster than the speed of light.

Any adequate account of laws must draw a distinction between the history of particular fact at a world and the laws of that world.¹⁴ For instance, it may be true that no one ever has or ever will read all of the *Oxford English Dictionary* (OED) aloud. But this proposition, even if true, is not a law. And any adequate account of laws must be able to make sense of the idea that worlds with different histories of particular fact share the same laws. And that's enough to provide us with the resources for saying the things we want to say about how the laws limit our causal powers.

We can define a set of propositions *L*, such that *L* is the set consisting of all and only the laws of our world, and we can define a set of worlds *W*, such that *W* includes all and only the worlds where the members of *L* (and no other propositions) are the laws. The members of *W* are the physically possible worlds (relative to our world).

With this in place, we can say how the laws constrain.

We have the power (ability, capacity, etc.) to do or cause *only* those things that are physically possible; that is, only those things that happen at some physically possible world – at one of the members of *W*.

So, for instance, since there is no physically possible world where anything accelerates to a speed faster than the speed of light, no human being is able to run faster than light. We lack this ability, and we lack it *because of the laws*. By contrast, there are physically possible worlds where someone reads all of the OED out loud. So even if it turns out that no living human being has the ability to do this, our having this ability is not ruled out by the laws.

But there are limits to how the laws constrain. The laws do not dictate all the details of the initial conditions of the universe, nor do they dictate all the details of how the world must be at any given time. This is so even at a deterministic world. If we claim that a person at a deterministic world has the ability, at a particular time, to do something other than what she actually does, we do not thereby contradict the uncontroversial claim, described above, about how the laws constrain. The incompatibilist needs an *argument* for the claim that at every deterministic world there is an *additional* constraint, and that this additional constraint robs us of the free will we would otherwise have, rendering us always unable to do otherwise.

There are, of course, such arguments, most notably the Consequence argument.¹⁵ I think, and have argued, that these arguments all fail.¹⁶ But that's not my present concern. In my experience, the uninitiated – my students – don't need to be given an argument for incompatibilism. Once they understand what the thesis of determinism says – and it does take them awhile to grasp that – they are *already* incompatibilists, of a particularly extreme variety.

My question, again, is, why do they respond this way? Why is compatibilism a belief that has to be acquired, by hard work? Shouldn't the truth be easier to believe? And even if I'm wrong, and the philosophical truth of the matter is that *free will* is incompatible with determinism, why do so many people respond to the thesis of determinism in the way that they do? Why do they think that the truth of determinism would mean that our commonsense beliefs about causation are all false?

4 Counterfactuals

I think that the explanation lies in certain counterfactuals that we entertain and believe true in situations that I call *choice situations*. Choice situations may be, but need not be, situations in which we have the *experience* of making a choice.

We have the experience of making a choice in situations where we don't immediately know what to do; situations in which we need to make up our mind by considering our options and choosing the one we think best, or right, or which most appeals. But choice situations also include situations in which we don't stop to think, either because the situation calls for a quick response or because we already know what to do. We learned, long ago, that touching a hot pot without oven mitts is a bad idea, that stopping at red lights is a good idea, and so on. I count these as choice situations because even if we did not act by first considering the alternatives and making a decision, it seems that we could have done so.

Some more examples: You are driving your car when a squirrel suddenly runs in the path of the car; to avoid hitting the squirrel, you could apply your brakes, swerve, or keep driving in the hope that the car will pass over the squirrel without hitting it. You have less than a second to decide. You keep driving. You walk into a room just in time to see a child reaching for a knife that has been left on a counter. Without pausing to think, you run over to stop the child.

Choice situations are ubiquitous. We've been in choice situations our entire lives, from the time we were babies, performing simple versions of scientific experiments in our attempts to find out how the world works.¹⁷ We learned about causation by being in choice situations.

What all choice situations have in common is that we believe that certain counterfactuals are true. When we deliberate, even briefly, for the purpose of deciding what to do, we begin our decision-making by asking certain kinds of counterfactual questions: What would happen if I did A? What would happen if I did B? We are good at evaluating these kinds of counterfactuals, and we think of them as objective in certain kinds of ways. We use them as the basis of our advice to others, and as the basis for some of our interventions with others. (We

run over to stop the child from grabbing the knife because we believe that if she got hold of it, she would probably hurt herself.)

Being good at counterfactual thinking is not the same as having a theory of counterfactuals. But if a theory of counterfactuals could be attributed to common sense, it might be the view that counterfactuals are like short stories.¹⁸ We evaluate the counterfactual *if A had been the case, then C would have been the case* by imagining a situation as much like the actual situation as possible given the truth of A, and we use our knowledge of how the world works, as well as our knowledge of particular facts, to figure out whether C would also be true. We don't worry *how* A got to be true; we just imagine that it is. Thinking this way works just fine so long as we are not thinking in terms of determinism or entire possible worlds, as opposed to small parts of worlds (the child and the knife, the squirrel and the car).

But in order to understand determinism, we must first understand the philosophical concept of a possible world, and we must understand the difference between a deterministic world and a non-deterministic world. When we understand this, we realize that our counterfactual short stories must be embedded within *entire possible worlds*. And this, I claim, is the source of our strong intuitive resistance to determinism. For it is natural to assume that our counterfactual short story can be "added" to the actual past and the actual laws. But this is possible only if the actual past and laws permit this counterfactual divergence from past history; that is, it is possible only if the actual world is *not deterministic*.

If determinism is true, there are *no* possible worlds that share our laws and our past. But we know that we make and criticize choices by using counterfactuals that we evaluate by holding fixed our knowledge of past facts as well as the laws, and we can't imagine doing it any other way. More generally, we evaluate counterfactuals about events like the child's dropping the knife or the squirrel's sudden dash onto the road by holding fixed the laws and the past, and we *can't think of any other way to do it*. So our first response to determinism – the first stage of the Three Stages of Denial – is to deny that it could possibly be true. This is a perfect instance of Hobart's "natural want of the analytic imagination."

The second stage – and first step in the direction of incompatibilism – is to understand determinism as "almost-determinism," making an exception for our own choices. This is a step backward from our earlier endorsement of No Exception, but it is a natural reaction. For experience tells us that we choose, and we choose between alternative actions A and B by evaluating counterfactuals about *what would happen if we chose A*, and *what would happen if we chose B*. So there must, it seems, be possible worlds with our past and our laws where we choose A and there must also be possible worlds with our past and laws where we choose B. But this doesn't help. We now reason as follows: "Okay, so determinism doesn't mean I have no free will. But determinism means that my choices never make any difference to the physical world. For if I chose otherwise, the laws and the past, *including the entire state of my brain and body*, would still be the same. And if my brain doesn't tell my body to move differently, it won't move differently." And thus the second stage is an epiphenomenal version of dualism. Our choices have no sufficient physical causes but they also have no physical effects.

We move to the third stage when we come to understand that determinism is a thesis without any exceptions; if determinism is true, then everything we do, including every *choice* we make, is constituted or realized by physical events and processes that instantiate strict deterministic laws. Our philosophy professor will try to explain that this doesn't preclude our choices from being *causes* as well as effects, and will try to explain some version

of compatibilism. But we will remain dismayed, for it will now seem that there are no non-vacuous counterfactuals. And it will seem that we cannot deliberate in the way that we are accustomed to doing. And while we might nod and agree that, of course, we still *make* choices, and our choices still cause our bodies to move in the ways appropriate to action, this won't mean what it used to mean. For causes are not what we thought they were, for they are no longer associated with non-vacuously true counterfactuals. We don't have the causal powers we thought we had; nor do horses, steam engines, or baseballs have the causal powers we thought they had. The causal power in the universe is located in the laws and the remote past, not in us. We will reluctantly conclude that determinism is true and free will is an illusion.

5 Determinism and Counterfactuals

Suppose that determinism is true. We still need a theory of counterfactuals. Even if we decide that the truth of determinism means the absence of free will, we still need to make decisions, and we make decisions by evaluating counterfactuals. And we still need to make causal judgments about a wide range of events not limited to the actions of human agents. And causal judgments typically go hand in hand with counterfactuals.

It remains an open scientific question whether determinism is true.¹⁹ And even if determinism turned out to be false in the way that libertarians hope, we cannot base our counterfactual thinking on this assumption: the counterfactuals that we believe true are not limited to counterfactuals about the choices of agents with free will. We need a theory of counterfactuals that is general enough to work at deterministic as well as non-deterministic worlds and general enough to apply to events like heart attacks, earthquakes, and plane crashes as well as the free choices of agents.

If determinism is true and *A* is any proposition that is actually false, there are no possible worlds with our laws and past history prior to *A* where *A* is true. We must give up at least one of:

Fixed Laws: If *A*, the laws would still be exactly the same as the actual laws.

Fixed Past: If *A*, the past prior to *A* would still have been exactly the same as the actual past.

Until David Lewis came along,²⁰ everyone who thought about counterfactuals assumed that we must retain Fixed Laws. The reigning assumption, in those days before possible worlds semantics, was that the counterfactual *if A, it would be that C* is true just in case *C* logically follows from *A* and some true premises. The notorious problem of cotenability, discovered by Goodman,²¹ was to say, without circularity and in a way that avoids counterfactual contradictions, which true statements are eligible to be included among the premises. But everyone assumed that the laws are always cotenable.

The philosophical orthodoxy has changed. David Lewis persuaded many that it is **Fixed Laws** that must go. It is now almost a platitude in certain circles that the correct way to evaluate counterfactuals, at least in so-called standard contexts, is by not permitting "back-tracking" counterfactuals;²² we must reason counterfactually in a way that preserves the past (or at least *most* of the past).

For a long time I was convinced by Lewis's theory and I defended Lewis against various charges.²³ I now think this a mistake. There is a better theory of counterfactuals – better in terms of what it tells us about counterfactuals and also better in terms of what it tells us about the relation between causes, counterfactuals, and laws.

6 How Lewis Persuaded Us: Counterfactual Knowledge and Time's Arrow

The starting point for thinking more rigorously about counterfactuals is the possible worlds semantics and logic developed independently by Stalnaker²⁴ and Lewis.²⁵ In this framework, the counterfactual *if A, it would be that C* is true iff C is true at the A-worlds most similar to the actual world.²⁶

Different theories of counterfactuals may be understood as offering different accounts of the relevant notion of similarity (or “closeness,” as it is sometimes called); these are, in effect, rival solutions to Goodman’s problem of cotenability.

In the book in which he introduced his logic for counterfactuals,²⁷ Lewis told us that we are to use our ordinary notion of *comparative overall similarity*, the one that we use when we say things like “Seattle is more like San Francisco than it is like Los Angeles.”²⁸ And he said that while similarity with respect to the laws is important, so is similarity with respect to matters of particular fact.

Kit Fine famously replied with the counterfactual *if Nixon had pressed the button, there would have been a nuclear holocaust*.²⁹ Since any world with a nuclear holocaust is vastly different from the world we fondly hope is our own, Lewis’s similarity metric appears to tell us that this counterfactual is false. But it is easy to stipulate that the facts are such that the counterfactual is true.

Lewis responded to Fine’s objection with his 1979 “Counterfactual Dependence and Time’s Arrow.”³⁰ The arguments that convinced many of us³¹ to give up **Fixed Laws** are mostly found there.³²

Lewis had two arguments – an epistemic argument and the time’s arrow argument.

The epistemic argument was that if we don’t evaluate counterfactuals by holding all, or at least most, of the past fixed (at the expense of the laws, if determinism is true), we won’t have any idea how to evaluate most of the counterfactuals we believe true.

For instance, consider Goodman’s example of a counterfactual truth we seem to know.³³ I’m holding a dry well-made match in a room with oxygen, no wind, and conditions favorable in every way. I don’t strike the match. If I had struck it, it would have lit. But if we evaluate this counterfactual by holding the laws rather than the past fixed, how can we say that it would light? After all, if determinism is true, then every world with our laws is a world where the past is different all the way back. At some of these worlds the struck match lights, but at others it fails to light because it is wet or defective or not surrounded by oxygen or something else goes wrong. On what basis do we say that some of these worlds are more similar to our world than others? So it seems that the most we are entitled to conclude is that if I had struck the match it *might* have lit and it *might not* have lit.

The key claim of the time’s arrow argument was that the counterfactuals that we ordinarily use are temporally asymmetric in a way that is not the product of stipulation but is in response to a pervasive but contingent fact about the world. Even though every detail of Lewis’s argument for this claim has been criticized,³⁴ the claim itself was striking and it seemed that there was something right about it. It was the combination of the purely negative epistemic argument against Fixed Law theories with the time’s arrow argument that did the trick. Even Bennett eventually gave up his Simple theory in favor of a theory that permits miracles.³⁵

The time’s arrow argument was complex.

First, Lewis claimed that counterfactuals are (with some qualifications)³⁶ temporally asymmetric. There are lots of true counterfactuals that say that if the present were different in some relatively specific way, the future would also be different in some relatively specific way. But specific backtracking counterfactuals are either false or not clearly true.³⁷

Second, Lewis said that the temporal asymmetry of counterfactuals corresponds to a theory that he called “Analysis 1.” Analysis 1 says that we (standardly) evaluate *if A, it would be that C* by considering worlds where the past is exactly like our past until a “transition period” shortly *before* the time of A and which obey our laws at all times *after* the time of A.³⁸

Third, Lewis argued that the temporal asymmetry of counterfactuals, as revealed by our use of Analysis 1, explains our belief that the past is in some sense “fixed” while the future is “open.” Since the temporal asymmetry of Analysis 1 is stipulated, it can’t be the last word.

Fourth, Lewis argued that Fine’s Nixon counterfactual is not a counterexample to overall similarity *properly understood*. It is a mistake, he said, to rely on our “offhand judgments” of overall similarity; rather, the correct similarity metric is whatever metric yields the right verdicts for uncontroversial counterfactuals like the one about Nixon.³⁹ He applied Analysis 1 to that case, and showed that it gets the right result. Same past; small miracle in Nixon’s brain causing Nixon to make the fateful decision and to push the button; events follow their lawful course and a nuclear holocaust ensues.

Fifth, Lewis set himself the task of seeing whether a *temporally neutral* similarity metric could, at a world like ours, pick out the same set of worlds as the temporally biased Analysis 1. Confining himself to the Nixon counterfactual, he succeeded.⁴⁰

Lewis’s temporally neutral similarity metric – I’ll call it “MaxiMin” – says that the most similar worlds are those which *maximize exact match of particular fact* taking into account not only the past but also the future; if it takes a *small miracle* to achieve a large spatiotemporal region of exact match of fact, so be it. Worlds with a small miracle are more similar than worlds with a *big miracle* (or many little miracles).

Lewis pointed out that only a *small miracle* (that extra neuron firing in Nixon’s brain) is required to set the counterfactual course of events in motion, but it takes a *huge miracle* (or many little miracles) to get the counterfactual sequence to converge back to the exact course of events at our world. The fatal signal must vanish on its way from the button to the rockets, Nixon’s fingerprints on the button must disappear, the click of the button on his tape-recorder must be erased, Nixon’s memories must be erased, and so on. Lewis said that the moral is that divergence from a world like ours⁴¹ is easier than perfect convergence to it. If determinism is true, then it takes a miracle to achieve either divergence or convergence, but “convergence takes very much more of a miracle.”⁴² The temporal asymmetry of counterfactuals is, he claimed, due to the fact that “the appropriate standards of similarity, themselves symmetric, respond to this asymmetry of miracles.”⁴³

Lewis made another explanatory claim: that the miracle asymmetry is explained by a *de facto* temporal asymmetry that he called “the asymmetry of overdetermination.” In a nutshell, his claim was that at worlds like ours events typically have only a single sufficient cause, but they leave many traces (the fingerprints, the click on the button, etc.), and it takes a miracle to break the link between an event and one of its traces; hence, the asymmetry of miracles.

Summing up, Lewis’s claim was that our use of the *temporally asymmetric* Analysis 1 is explained by our use of the *temporally symmetric* MaxiMin together with a contingent fact about our world – the overdetermination asymmetry.

7 Why We Should Not Be Persuaded

Lewis had just one argument against Fixed Law theories – the epistemic argument, which says that if we hold the laws fixed, we won’t have any idea how to evaluate any counterfactuals, including the uncontroversially true ones, like the Nixon and match counterfactuals. By contrast, Analysis 1 makes it easy to evaluate these counterfactuals.

If we find this argument persuasive, it is because we are assuming, with Lewis, that the correct similarity metric is one that compares worlds with respect to *similarity during all time* – past, present, and future. If determinism is true, this means that a Fixed Law theory tells us to compare worlds with different histories all the way back to the beginning of time. And if that's the case, it's hard to see how a Fixed Law theory can tell us which worlds are most like our own.

The epistemic argument fails against Bennett's version of a Fixed Law theory, his Simple theory.⁴⁴ The Simple theory says that the most similar worlds are worlds with the *same laws* as our world and which are most like our world with respect to particular fact at the *time of the antecedent*. About the match, it says that the relevant worlds are ones where the match and its surroundings are almost exactly the way they actually are – the match is dry, well-made, in a room with oxygen and without wind or rain, and so on. This gives us the right results; at these worlds, my striking the match is followed by its lighting. About Nixon, the Simple theory says that the relevant worlds are ones where everything at the time of Nixon's button-pressing is almost exactly the way it actually was; the Doomsday machine is in perfect working order, as is Nixon's tape recorder, and so on. More generally, since we know about the past by knowing its many traces in the present, the Simple theory provides us with the same epistemic access to many counterfactuals as Analysis 1.

The time's arrow argument was based on a claim that seems right – that there is a striking temporal asymmetry in the counterfactuals we think true.

Reflection on this temporal asymmetry might seem to support:

Fixed Past: If the present were different in any way, the past would (still) be exactly as it actually was.

But the counterfactual data that Lewis produced as evidence for the temporal asymmetry of counterfactuals supports only:

Indeterminate Past: If the present were different in any way, the past would be different, but not in any very specific way.

And *Indeterminate Past* is consistent with Fixed Law theories of counterfactuals, including the Simple theory.

Contrary to what is sometimes claimed (and contrary to what he sometimes says), Lewis does not hold *Fixed Past*. Lewis says that in order for there to be a “smooth” and “orderly” transition from the actual past to the counterfactual future, some part of the past (prior to the antecedent) must be different.⁴⁵ But, he reassures us, the past would not be different in any very specific way.⁴⁶

Lewis's Analysis 1 thus requires us to think of the past as *Fixed* to a certain point in time and *Indeterminate* thereafter. He suggests that it is only the “immediate past” that is *Indeterminate*, but he cannot mean this. The length of the transition period will vary, depending on the counterfactual. We can argue about whether Kennedy would have won if he had run for a second term as president (in 1964) But in arguing this, we consider worlds where Kennedy was not assassinated (in 1963).

But note where we are now. We began with Lewis's initial and plausible observation that counterfactuals are temporally asymmetric. Lewis moved swiftly from there to the claim that this asymmetry explains our time's arrow belief and is in turn explained by the miracle-permitting **MaxiMin** together with the overdetermination asymmetry. But the temporal asymmetry of counterfactuals explains our time's arrow belief only if *Fixed Past* is true, and *Fixed Past* is not true, by Lewis's own admission. And the counterfactual data cited by Lewis doesn't give us reason to prefer a miracle-permitting theory to a Fixed Law theory.

Lewis had only one argument against Fixed Law theories – the epistemic argument. But that argument fails against the Simple theory. The Simple theory generates the correct verdicts for the uncontroversially true counterfactuals, like the match and Nixon counterfactual, and therefore passes Lewis’s own test for how we should think about similarity. (“we must use what we know about counterfactuals to find the appropriate similarity relation, not the other way around”).⁴⁷ In the absence of a reason to reject the Simple theory, we have no reason to accept Lewis’s claim that we evaluate counterfactuals according to a theory that permits miracles.

8 Some Reasons to Prefer the Simple Theory

The first reason is that it is simple, unlike either Analysis 1 or MaxiMin which require us to balance different similarity respects – smallness of miracle against size of region of exact match of particular fact. The Simple theory says that the closest worlds are worlds with exactly our laws – no messing about with miracles – and which are otherwise most like our world *at the time of the antecedent*. This makes the Simple theory a more plausible candidate as an explanation for how we actually learned to use counterfactuals, from the time we were babies, to figure out the causal facts about the objects around us.

Objection: Didn’t you say that we evaluate counterfactuals by holding fixed the laws *and* the past? Wasn’t this your explanation of the Three Stages of Denial?

Reply: No, I was careful not to say this. I said that we are – and have been from an early age – competent users of counterfactuals without having any articulated theory of *how* we do this. My claim was that Three Stages of Denial are explained by what happens in the philosophy classroom when we are first introduced to determinism and forced to take seriously the thought that it might, for all we know, be the truth about our world. We respond by articulating, for the first time, a theory of counterfactuals. I did not say it then, but I will say it now: this theory isn’t just a bad theory of counterfactuals; it’s not a theory we ever used.⁴⁸ We make a mistake when we attribute this theory to ourselves, but it’s not a crazy mistake, given that the counterfactuals we are most familiar with – choice counterfactuals and counterfactuals about alternative behaviors of nearby agents and objects – can be evaluated while *ignoring* the past. But ignoring the past is not the same as holding it fixed. And holding some past facts counterfactually fixed is not the same as holding the entire past fixed. The fact that we can also understand and evaluate backtracking counterfactuals shows that we don’t think that *all of past history* is counterfactually fixed.

Taking determinism seriously means facing up to the fact that we cannot evaluate counterfactuals by limiting our attention to worlds with the same laws and same past. Something’s got to give. The correct theory of counterfactuals *must* be a theory that gives up either Fixed Past or Fixed Laws. The Simple theory says that it is Fixed Past that must yield; Lewis argued that it is Fixed Laws that must go. (And, though he did not advertise it, he also rejected Fixed Past.) Both theories are able, with some adjustments,⁴⁹ to account for the counterfactual data – the forwards counterfactuals that are as clearly and uncontroversially true as counterfactuals ever are. But the Simple theory has the theoretic virtue of simplicity, and because of this it is more plausible that it is the theory that we actually use.

The second reason for preferring the Simple theory is that its similarity metric is temporally neutral, permitting counterfactual reasoning in both temporal directions. It is therefore able to accommodate the plausible claim with which Lewis began, about the way in which counterfactuals are temporally asymmetric, without requiring it, as Analysis 1 does. And

because the Simple theory doesn't stipulate a temporal asymmetry, it does not need to provide an explanation for why we evaluate counterfactuals in this temporally asymmetric way. It leaves it open, in a way that Lewis's theory does not, what the correct explanation of this temporal asymmetry is. It thus has the theoretic virtue of modesty, of assuming no more than is required to explain the counterfactual data.

The third, and most important, reason for preferring the Simple theory is that it takes the laws seriously – by holding them counterfactually fixed. This should be our default assumption, when looking for a theory of counterfactuals. We shouldn't accept miracles unless we have to. There is a tight link between laws and counterfactuals; the laws tell us not only what actually happens but what *would happen even if things were different*. When we believe that something is a law, we believe that it *would hold no matter what else happens, no matter what we do or try to do*. When we consider counterfactuals for the purpose of making decisions about what to do (and not to do – better not step off that balcony without a safety net!), we rely on our knowledge of the laws.

The Simple theory accounts for these commonsense beliefs in a simple and straightforward way: by telling us that the worlds most similar to ours are always (except for the special case of counterlegals) worlds with *exactly* our laws. Lewis can account for our commonsense beliefs as well, but only in a more complicated way, and only up to a point. He is able to account for the role of laws in our counterfactual thinking so far as the future is concerned. And, depending on your view of the nature of laws, he may be able to dismiss worries about how to think about the laws at his small miracle worlds.⁵⁰ But he isn't able to endorse the commonsense platitude that the laws – *all the laws* – would hold no matter what happens. And he has to draw some fine distinctions to avoid having to say that the ability of a free deterministic agent to act otherwise is not an incredible ability to break the laws.⁵¹

It is sometimes thought that Lewis's theory of counterfactuals is a consequence of his Humean view of laws, and it has even been said that a compatibilist who is a Humean about the laws must be a "local miracles compatibilist." This is a mistake. It is true that Lewis's view of the laws was one of the factors that led him to his theory of counterfactuals. But it is by no means mandatory for a Humean to accept Lewis's theory of counterfactuals. It is worth remembering that until the time of Lewis's *Counterfactuals* (1973a), everyone assumed that the laws are always cotenable with the antecedent of a counterfactual – that is, that the laws are counterfactually fixed. And in those days almost everyone held the Humean view that laws are a subset of the true universal generalizations.

9 Concluding Remarks

What does all this have to do with the free will/determinism problem?

At least this. To understand the truth about free will and determinism, we need to understand counterfactuals. I've argued that our beliefs about free will and our causal powers are closely connected with our beliefs about certain kinds of counterfactuals – the counterfactuals that we believe true in choice situations – and I've argued that there is an apparent tension between these counterfactuals and the truth of determinism. When we engage in counterfactual reasoning in choice situations, it seems that we hold *both* the laws and the past counterfactually fixed. But if determinism is true, then either the past is not counterfactually fixed or the laws are not counterfactually fixed. We are faced, it seems, with a hard choice. Lewis persuaded many that the lesser evil is to keep the past (or at least most, or much, or some, of the past) fixed, at the cost of a small miracle. But this, I have argued,

is a mistake. In arguing this, I have not appealed to my belief that free will is compatible with determinism or even my belief that the existence of ordinary sorts of causal powers (of baseballs, horses, and steam engines) is compatible with determinism. I have appealed only to the kinds of considerations that count in favor of a theory of counterfactuals. I've argued that the Simple theory is a better theory of counterfactuals. If it is the better theory of counterfactuals, then it is the one we should use when we think and argue about the free will/determinism problem.

As I said, I do not claim that the Simple theory makes it easier for us to defend compatibilism. But I do claim that the Simple theory provides us with the resources for thinking more clearly about the free will/determinism problem. In order to think clearly about the problem, we must take determinism seriously. Determinism isn't just the claim that there are laws of nature that describe general truths about the way the world *is*; it is the claim that the laws, somehow, circumscribe the ways the world *can* be.

Humeans are often accused of not acknowledging this, but that's not fair. Here is Lewis's definition of determinism:

A deterministic system of laws is one such that, whenever two possible worlds both obey the laws perfectly, then either they are exactly alike throughout all of time, or else they are not exactly alike through any stretch of time. They are alike always or never. They do not diverge, matching perfectly in their initial segments but not thereafter; neither do they converge.⁵²

By explicitly quantifying over nomologically possible worlds, Lewis captures the modal commitments of determinism. His definition makes clear how *much* we are assuming when we assume for the sake of any philosophical argument that determinism is true.

Given this understanding of the modal commitments of determinism, we might expect Lewis to agree that there are counterfactual commitments as well. If determinism, as Lewis understands it, is true, it seems reasonable to suppose that Counterfactual Determinism is true as well.

Counterfactual Determinism: If the world had been different in any respect at any time, then it would have been different *in some respect* at every moment in the future and the past.

But Lewis denies Counterfactual Determinism. He agrees that if the world (assuming determinism) were different now, it would also be different in some respect at every future time. But he denies that if the world (assuming determinism) were different now, it would also be different in some respect at every past time.

The Simple theory, by contrast, affirms Counterfactual Determinism. As Bennett says, "It seems to me just plainly true that if the actual world is deterministic then if a certain pebble had rolled at a moment when in fact it did not roll, the entire previous history of the world would have had to be different."⁵³

In affirming that determinism entails Counterfactual Determinism, the Simple theory provides us with a unified framework for addressing questions about how, if at all, we can have the free will we think we have, and the causal powers we think we have if determinism turns out to be the truth about our world. More generally, it provides a unified framework for addressing more fundamental questions about how steam engines, horses, and baseballs can have the kind of causal powers we think they have if determinism is true. And it is only by having answers to these hard questions that we can hope to provide a compatibilist solution to the free will/determinism problem in a way that avoids the usual charges of "wretched subterfuge" and "quagmire of evasion."⁵⁴

Notes

- 1 In this essay, I use “compatibilist” to refer to someone who believes that the truth of determinism is compatible with our having the free will we think we have, including the ability to do otherwise. I use “incompatibilist” to refer to someone who believes that the truth of determinism is incompatible with our having the free will we think we have, including the ability to do otherwise.
- 2 Hobart 1934.
- 3 Lewis 1979.
- 4 Bennett 1984.
- 5 Gopnik, Meltzoff, and Kuhl 1990, Chapter 3, “What Children Learn about Things,” especially “Making Things Happen” and “Kinds of Things,” pp. 73–82.
- 6 What I have just argued should be understood as an empirical claim about how we in fact (typically) acquire our causal beliefs. I’m not claiming that there is no other way to acquire causal beliefs. For discussion of some empirical research on the causal counterfactual reasoning abilities of young children, see German and Nichols 2003.
- 7 For more detail, see Vihvelin 2013, Chapter 3.
- 8 Chisholm 1964.
- 9 See Vihvelin 2013 for argument that the possession of these agential powers is compatible with determinism as well as indeterminism and that they may be understood as dispositions or bundles of dispositions, differing in complexity but not in kind from the dispositions of natural objects like horses, steam engines, rocks, and lumps of sugar.
- 10 Van Inwagen 1983; Beebe and Mele 2002.
- 11 Lewis 1979; Earman 1986.
- 12 Vihvelin and Tomkow 2016.
- 13 For more on determinism, and for why determinism should not be confused with various claims about causation (or physicalism or naturalism), see Vihvelin 2013, 2017.
- 14 I’m not claiming that every Humean theory of laws satisfies this (and other) conditions of adequacy for lawhood. See, for instance, Armstrong 1983 for a critique of a Humean theory that he calls “the Naïve theory.”
- 15 van Inwagen 1983.
- 16 Vihvelin 2013.
- 17 For an excellent account, by three leading psychologists, of some research about the “scientist in the crib,” see Gopnik, Meltzoff, and Kuhl 1999.
- 18 Kim and Maslen 2006.
- 19 Loewer 1996.
- 20 Lewis 1973a, 1973b, 1979.
- 21 Goodman 1947.
- 22 Collins, Hall, and Paul 2004.
- 23 Vihvelin 2000, 2013.
- 24 Stalnaker 1968, 1984.
- 25 Lewis 1973a.
- 26 This is rough, since Lewis rejects the Limit assumption – that there is a closest set of worlds. His official account says that $A > C$ is non-vacuously true iff there is some world where A and C are both true that is more similar to our world than any world where A is true and C is false.
- 27 Ibid.
- 28 Ibid., p. 92. “Somehow we do have a familiar notion of comparative overall similarity, even of comparative similarity of big, complicated things like whole people, whole cities, or even – I think – whole possible worlds.”
- 29 Fine 1975.
- 30 Lewis 1979.
- 31 I was among those convinced and I even used Lewis’s theory to defend a bold kind of compatibilism I called “Libertarian Compatibilism” (Vihvelin 2000). In brief, my claim was that even if

- determinism is true, we are sometimes able to choose otherwise, given the actual past up to the time of our choice; our choice is the small miracle. I now think that my claims in that paper were mistaken. This wasn't entirely because I was relying on Lewis's theory of counterfactuals, but that was part of it. For discussion of the question of whether a "local miracle compatibilist" must say that a free agent's choice is the divergence miracle, see Beebe 2003 and Graham 2008.
- 32 Though not entirely. The groundwork was laid in Lewis 1973a, especially in Chapter 3 ("Comparisons"), in which he discussed the problem of cotenability) and Chapter 4 ("Foundations"), in which he argued that the "limited vagueness" of overall similarity of worlds "accounts nicely for the limited vagueness" of counterfactuals."
- 33 Goodman 1947.
- 34 Bennett 1984; Horwich 1987; Goggans 1992; Price 1992; Vihvelin 1995; Elga 2000; Tomkow 2013; Mackie 2014; Wilson 2014; Vihvelin and Tomkow 2015; Dorr 2016; Vihvelin and Tomkow 2016.
- 35 Bennett 2003, Chapter 13. For a close examination and critique of Bennett's rejection of his earlier theory, see Tomkow 2013. Bennett 2003 rejects his earlier Simple theory, but falls short of endorsing a Fixed Past theory. The theory that he endorses appears to be neutral between three different ways of spelling out the details of what he calls "forks" – indeterministic forks, small miracle forks, and "exploding difference" forks. An "exploding difference" fork takes place when two worlds with the same deterministic laws are "invisibly unlike" but "alike in every respect we would ever notice or think about" for "aeons" and then suddenly, legally and improbably, diverge. An "exploding difference" theory of counterfactuals is a Fixed Law theory.
- 36 He said that the asymmetry might not hold in a time machine, or at the edge of a black hole, or at a possible world consisting of a solitary atom in the void. And he granted that it does not hold in what he called "non-standard" contexts (Lewis 1979, pp. 33–34).
- 37 His example was of typing words on a page. "Suppose I were typing different words. Then plainly tomorrow would be different also. For instance, different words would appear on the page. Would yesterday also be different? Invited to answer, you will perhaps come up with something. But I do not think there is anything you can say about how yesterday would be that will seem clearly and uncontroversially true" (ibid., p. 32).
- 38 Ibid., pp. 39–40.
- 39 Ibid., pp. 42–43.
- 40 For critique, see Elga 2000.
- 41 By "world like ours" he means worlds where counterfactuals are temporally asymmetric in the way they actually are.
- 42 Lewis 1979, p. 49.
- 43 Lewis 1979, p. 49.
- 44 Bennett 1984.
- 45 He uses the example of a stationary match a foot away from the nearest striking surface at time t . "If the match had been struck at t , would it have travelled a foot in no time at all? No, we should sacrifice the independence of the immediate past to provide an orderly transition from actual past to counterfactual present and future" (ibid., p. 40).
- 46 ibid., p. 40.
- 47 ibid., p. 43.
- 48 Thanks to Helen Beebe for pressing me on this point.
- 49 See Bennett 2003, pp. 219–221 and pp. 234–240 for discussion of counterexamples and problem cases, and for suggested remedies.
- 50 "The violated deterministic law has presumably not been replaced by a contrary law. Indeed, a version of the violated law, complicated and weakened by a clause to permit the one exception, may still be simple and strong enough to survive as a law" (Lewis 1973a, p. 75).
- 51 Lewis 1981 distinguishes a Strong ability to perform actions which would either be or cause law-breaking events from a Weak ability to act so that some law would have been broken, and claims that his compatibilism commits him only to the claim that a free deterministic agent has the Weak ability. See Beebe 2003 for critique; see Graham 2008 for defense of Lewis.

- 52 Lewis 1979, p. 37.
- 53 Bennett 1984, p. 68.
- 54 Ancestors of this paper were presented at the Free Will and Causation conference at the Collège de France in Paris in September 2016 and at the Free Will and the Ability to Do Otherwise workshop at the University of Belval in Luxembourg in June 2017. I am grateful to the organizers of both conferences – to Claudine Tiercelin and Jean-Baptiste Guillon at the Collège de France and to Frank Hoffman and Christian Loew at the University of Belval – and to all who participated for their helpful comments. Thanks also to an anonymous referee for Wiley-Blackwell and to Joe Campbell, for extensive written comments on a draft of this paper. Finally, I am grateful to Terrance Tomkow, for hours of extremely valuable discussion about the counterfactuals part of the paper.

Bibliography

- Armstrong, D. (1983). *What Is a Law of Nature?* Cambridge: Cambridge University Press.
- Beebe, H. (2003). Local miracle compatibilism. *Nous* 37: 258–277.
- Beebe, H. and Mele, A. (2002). Humean compatibilism. *Mind* 111 (442): 201–223.
- Bennett, J. (1984). Counterfactuals and temporal direction. *Philosophical Review* 93: 57–91.
- Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.
- Chisholm, R. (1964). Human freedom and the self. *The Lindley Lectures*, Department of Philosophy, University of Kansas.
- Collins, J., Hall, N., and Paul, L.A. (eds.) (2004). *Causation and Counterfactuals*. Cambridge, MA: MIT Press, A Bradford Book.
- Dorr, C. (2016). Against counterfactual miracles. *Philosophical Review* 125 (2): 241–286.
- Earman, J. (1986). *A Primer on Determinism*. Dordrecht, Holland: D. Reidel Publishing.
- Elga, A. (2000). Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science* 68 (Supp. vol.1): 313–324.
- Fine, K. (1975). Review of counterfactuals. *Mind* 84: 451–458.
- German, T. and Nichols, S. (2003). Children's counterfactual inferences about long and short causal chains. *Developmental Science* 6 (5): 514–523.
- Goggans, P. (1992). Do the closest counterfactual worlds contain miracles? *Pacific Philosophical Quarterly* 73: 137–149.
- Goodman, N. (1947). The problem of counterfactual conditionals. *The Journal of Philosophy* 44: 113–128.
- Gopnik, A., Meltzoff, A.N., and Kuhl, P.K. (1999). *The Scientist in the Crib: Minds, Brains, and How Children Learn*. New York: William Morrow and Company, Inc.
- Graham, P. (2008). A defense of local miracle compatibilism. *Philosophical Studies* 140 (1): 65–82.
- Hobart, R.E. (1934). Free will as involving determination and inconceivable without it. *Mind* 63: 1–27.
- Horwich, P. (1987). *Asymmetries in Time*. Cambridge, MA: MIT Press.
- Kim, S. and Maslen, C. (2006). Counterfactuals as short stories. *Philosophical Studies* 129 (1): 81–117.
- Lewis, D. (1973a). *Counterfactuals*. Cambridge, MA: Harvard University Press.
- Lewis, D. (1973b). Causation. *The Journal of Philosophy* 70: 556–567.
- Lewis, D. (1979). Counterfactual dependence and time's arrow. *Nous* 13: 455–476.
- Lewis, D. (1981). Are we free to break the laws? *Theoria* 47: 113–121.
- Loewer, B. (1996). Freedom from physics: quantum mechanics and free will. *Philosophical Topics* 24: 91–112.
- Mackie, P. (2014). Counterfactuals and the fixity of the past. *Philosophical Studies* 168: 397–415.
- Price, H. (1992). Agency and causal asymmetry. *Mind* 101 (403): 501–520.
- Stalnaker, R. (1968). A theory of conditionals. In: *Studies in Logical Theory* (ed. N. Rescher), 98–112. Oxford: Blackwell.
- Stalnaker, R. (1984). *Inquiry*. Cambridge, MA: Bradford Books.
- Tomkow, T. (2013). The simple theory of counterfactuals. www.tomkow.com. Accessed on 9th July, 2013.

- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.
- Vihvelin, K. (1995). Causes, effects, and counterfactual dependence. *Australasian Journal of Philosophy* 73: 560–573.
- Vihvelin, K. (2000). Libertarian compatibilism. In: *Action and Freedom, Philosophical Perspectives* (ed. J. Tomberlin), Vol. 14, 139–166. Malden, MA: Blackwell.
- Vihvelin, K. (2013). *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. New York: Oxford University Press.
- Vihvelin, K. (2017). Arguments for incompatibilism. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu>. Accessed on 9th July, 2013.
- Vihvelin, K. and Tomkow, T. (2015). Counterfactuals: the short course. www.vihvelin.com. Accessed on 20th August, 2015.
- Vihvelin, K. and Tomkow, T. (2016). Determinism. www.vihvelin.com. Accessed on 30th July, 2016.
- Wilson, J. (2014). Hume's dictum and the asymmetry of counterfactual dependence. In: *Chance and Temporal Asymmetry* (ed. A. Wilson). Oxford: Oxford University Press.

Free Will and Implicit Attitudes

NEIL LEVY AND JESSICA WRIGHT

Disputes over free will and moral responsibility, as we know them today, are the product of a continuous debate that dates back at least as far as Aristotle. The control and epistemic conditions that many philosophers identify as necessary and sufficient conditions for freedom of action are direct outgrowths of the conditions he identified in the *Nicomachean Ethics* (Fischer and Ravizza 1998, pp. 12–13). In some ways, the continuity of these debates from antiquity to the present should give us pause. After all, there is a gulf in our understanding of the world that separates us from the Greeks. The scientific revolution cast Aristotelian physics and biology in the dustbin of history. Our medicine is not continuous with his; why should our theories of agency be?

Theories of agency are, after all, in part theories about *us*: about particular, evolved, embodied beings. Perhaps – *perhaps* – debates concerning the metaphysics of free will can be conducted without reference to the kinds of animals we actually are.¹ But when it comes to assessing the extent to which we act freely, in the actual world, we need to be guided by the best available evidence concerning the springs of our action. Our naïve intuitions about the kinds of creatures we are may be off-track. Contemporary cognitive science seems to reveal that we are beings that are in some ways very different than we had thought. It is past time that our theories of freedom reflected how we actually are, rather than how we imagine ourselves to be.

In this chapter, we will survey some of the ways in which philosophers have responded to this challenge in recent years, with a special focus on implicit attitudes. The nature and function of these attitudes is an extremely significant, and fascinating, topic in its own right. The literature on it is small, but high quality (and growing rapidly). It raises a number of concerns about our agency and about whether we are responsible for our actions, concerns that generalize to other challenges from the cognitive sciences. It is therefore deserving of sustained attention. In this chapter we will begin with a critical discussion of the empirical studies on which arguments about the nature of implicit attitudes are based; then we will turn to a discussion of issues in the interpretation of those empirical results. We then turn to a discussion of the nature of individual responsibility, and draw out some of the consequences of the debate over implicit cognition for theories of moral responsibility.

1 The Nature of Implicit Attitudes

It is not possible to give a definition of implicit attitudes that is uncontroversial: their nature is a central bone of contention in the scientific and philosophical literature. Rather than define them, therefore, we will attempt to fix the reference of the phrase, and focus on cases of central importance to our topic. When people are asked their opinions about important moral issues, they report their *explicit* attitudes. These are the attitudes to which they are committed; usually, they are the attitudes they take to be supported by the facts.² Thus, for instance, asked what we believe about the intelligence of women, we normally employ the transparency method famously outlined by Gareth Evans (1982), looking to the world to answer the question and attribute a belief to ourselves. Most of us say “I believe that women are every bit as intelligent as men.” But we may have implicit attitudes that tell a different story about us (Storage et al. 2020 report evidence that “genius” and “brilliance” are associated more with men than with woman).

Implicit attitudes may be measured in a variety of ways. For reasons of space, we outline just one. The *Implicit Association Test* (IAT) is by the most prevalent measure of implicit bias: literally millions of people have taken the test since it was developed in the mid 1990s. The IAT is designed to measure the strength of associations between concepts, by measuring the reaction times of participants. IATs may be used to measure test subjects’ attitudes about any social category we like: race, sexuality, gender, and so on. For example, IATs are often used to assess the strength of the association between racial categories and negative and positive concepts. Participants are asked to react by pressing a key on a computer keyboard whenever an image or a word corresponding to the category assigned to the key is flashed up on the screen, while pressing another key whenever a different category is flashed up. So, for instance, photographs of Black and White people may flash up, interspersed with words describing good (“flower”; “puppy”; “joy”; “sunshine”) or bad (“death”; “sadness”; “cancer”; “ruin”) concepts and things. When the stimulus is “Black” or “good,” the participant is to press key one, whereas when the stimulus is “White” or “bad,” the participant is to press key two. Assignments alter from block to block, so that some blocks participants get White/good and Black/bad and others White/bad and Black/good. The oft-replicated finding is that the great majority of White participants, and a significant minority of Black participants are quicker to respond when the pairing is White/good and Black/bad than vice-versa (Greenwald, McGhee and Schwartz 1998; Nosek, Greenwald and Banaji 2007; Greenwald et al. 2009). This has been taken to be evidence that we tend to *associate* being Black with being bad (perhaps as a consequence of our repeated exposure to cultural stereotypes).

If the IAT measures associations, then on most theories of belief, it does not measure beliefs. This is because beliefs are typically taken to be attitudes with propositional content; loosely: content reportable by a whole and declarative sentence. Associations are attitudes that do not contain contents that are best described as propositional, instead referring to loose relationships (of inhibition or activation) between concepts (e.g., “salt” brings to mind “pepper”).

Why should it matter whether or not we associate *Black* with *bad*? One possibility is that associations are constituents of beliefs: if we are dispositionalists about beliefs, holding that to believe that *p* is to be disposed to judge, act and feel in *p*-appropriate ways, we might think that being disposed to associate *Black* with *bad* is to believe that *Black people are bad* (see Schwitzgebel 2001, 2010: note, though, that Schwitzgebel does not think that the implicitly biased believe the content of their biases, rather he thinks that their dispositions show that they are in-between believers, neither believing the content nor disbelieving it). More significantly,

implicit attitudes can cause behaviors that conflict with our explicit, self-reported attitudes in a phenomenon called attitude dissociation.³ If the IAT measures associations, then this also explains the phenomenon of attitude dissociation: implicit and explicit attitudes dissociate because they are different *kinds* of attitudes. Moreover, because explicit attitudes are measured by self-report but implicit attitudes are measured via other behavioral cues, it is safe to assume that the attitudes one takes oneself to have are the attitudes that one asserts or endorses (explicit attitudes), while one's implicit attitudes are (in these cases) unendorsed by or unknown to the agent.

Implicit attitudes are thought to manifest in our behavior in a number of ways and across many different domains. For example, controlling for the influence of explicit prejudice on behavior, implicit bias against Black people predicted an unwillingness to vote for Barack Obama in the 2008 election (Payne et al. 2010). Implicit bias can also predict the future voting behavior of presently undecided voters (again controlling for one's implicit attitudes) (Lundberg and Payne 2014). There are well-supported theoretical models of cognition that explain how implicit attitudes affect behavior, and which imply that we should find these effects occurring across many domains. On standard models of cognition, explicit attitudes correspond (roughly) to the domain of type-two cognition. Type-two cognition is slow, effortful, demanding of resources, typically conscious, serial and domain-general. Implicit biases, however, correspond (again roughly) to the domain of type-one cognition, which is fast, effortless, typically unconscious and domain-specific (Evans 2008). Conscious, type-two, reasoning is powerful, but is unable to operate without pervasive guidance from type-one cognition (we would face an intractable combinatorial explosion of possibilities if we did not have implicit processes that narrow down the space of options we take seriously, for example). Type-one processing assigns weight to options, and may cause behavior quite directly under some conditions (when we must respond rapidly, for example). That predicts that under many conditions we will act in ways that are modulated by our implicit attitudes (when we are unable to detect how they affect our options, for example) and that under some conditions these attitudes will be very significant determinants of our behavior (Greenwald et al. 2009).

The predictive validity of implicit attitudes – roughly, what proportion of our behavior they can explain – is highly contentious (Oswald et al. 2013; Greenwald, Banaji and Nosek 2015). It is not controversial that our explicit attitudes explain and predict the majority of our behavior. Nevertheless, there are good reasons to think that for very many of us, some of our behaviors will have contents that they would not have had were our explicit attitudes able to control our behavior without the influence of our implicit attitudes. Some of these behaviors will be consequential. An implicit association of “genius” more with men than with women, for example, can be predicted to have effects on hiring decisions for jobs that require high intelligence (like “philosophy professor”). While the influence of implicit attitudes on behavior is typically very small, even under conditions in which we have the opportunity for reflection small influences can be decisive. When we are assessing a shortlist of candidates for a job, tie breaks between more or less well-qualified applicants are routine. How do we rank candidate A, who has two publications in top 5 journals, against B, who has 6 publications in top 10 journals though none in top 5 journals? In these conditions, our implicit biases may be decisive, causing us to confabulate criteria of merit (see Uhlmann and Cohen 2005).

Many people find it unsettling to learn that some of their behavior is caused by implicit attitudes (Howell et al. 2013), but those that are dissociated from our endorsed attitudes pose a particular problem for our theories of freedom and responsibility. In these cases, agents perform an action that they would not have, were only their endorsed attitudes in play. Are

we responsible for such actions? Are we blameworthy for them? These are questions that have sparked a lively debate. On the near universal assumption that we can only be directly responsible for actions if we perform them freely, these are debates over whether implicit attitudes undermine or reduce our freedom.

2 Implicit Attitudes and Freedom

It is important to distinguish the set of issues we are concerned with here from nearby questions. There are interesting and important questions to be asked about our moral responsibility for possession of implicit biases and implicit attitudes more generally. If, as many people believe, we acquire these biases as children, as a result of a certain kind of enculturation, there seems to be a *prima facie* case for excusing us blame for possessing them (see Holroyd 2012 for critical discussion). There are interesting and important questions to be asked about our forward-looking responsibility: our responsibility to reduce our biases, if we can, and to change the social structures that make their acquisition so likely. Here we focus on responsibility in the backwards-looking, or basic desert (Pereboom 2013) sense: do we satisfy the freedom-relevant considerations on moral responsibility with regard to the actions themselves?

It is standard to approach questions like this by utilizing the method of reflective equilibrium. Utilizing this method involves working from two ends simultaneously: putting our theories to the test of the intuitions generated by consideration of cases and our intuitions to the test of our theories. Insofar as we take intuitions about cases as data, we have grounds for revising our theories when they yield the “wrong” results. There seems to be *prima facie* grounds for excusing agents responsibility for actions caused by their implicit biases. They seem not to satisfy epistemic conditions with regard to these actions, for instance, on the plausible grounds that satisfaction of these conditions required awareness of the moral character of our actions (Levy 2014; see Washington and Kelly 2016 for a dissenting view). However, this seems an unacceptable result to many philosophers who have considered the cases. Utilizing the standard methodology, they therefore hold that we must revise our accounts of moral responsibility such that agents count as responsible for these actions (Faucher 2016).

However, there are grounds for doubting that the standard methodology is appropriate in this context. A great deal depends, here, on our grounds for holding that intuitions give us insight into the nature of our concepts or their application conditions. We assume a broadly realist picture of agency and of freedom; that is, we assume that facts about the world and the kind of agents humans really are constrain theories. Theories may be true or false, depending on how closely they reflect these facts. That being the case, our intuitions may be off-track in certain contexts. Suppose, as we think is plausible, that our intuitions are well designed to track the conditions that make agents morally responsible for their actions when their actions are caused by the kinds of states folk psychology recognizes (beliefs, desires, intentions, and so on). In that case, they may generate the wrong results when behaviors are importantly influenced by states of a kind that go unrecognized in folk psychology. We may have the intuition that an agent is responsible for a token action because the overt features of that action trigger domain-specific mechanisms designed to track beliefs and desires that the agent does not actually possess. If that’s the case, we would have good reason to set aside these intuitions.⁴

If this line of thought is on the right track, we need to understand the nature of implicit attitudes in order to assess our moral responsibility for the actions they help to cause. If implicit attitudes are a kind of state recognized by folk psychology (or that the relevant mechanisms are designed to track), then reliance on our single case intuitions may be appropriate.

If they are not, then, we must take a different approach: asking instead whether agents who act on them satisfy the conditions laid down by our best *theories* of moral responsibility.

Above, we mentioned that the IAT was widely held to measure the strength of our associations between concepts. Associations may well be among the kinds of states that folk psychology recognizes. But they may not be among the kinds of states that our intuition-generating mechanisms are designed to track. If implicit attitudes are mere associations, then a disposition to blame agents for the actions they help cause may be unreliable.

However, on some accounts implicit attitudes are not mere associations. Surveying a wide range of data concerning how implicit states cause actions and how they interact with other representational states, Mandelbaum (2016) argues that implicit attitudes belong to the same psychological class as beliefs. Since beliefs are certainly among the kind of states that our intuition-generating mechanisms respond to, Mandelbaum's view suggests that we ought to take our intuitions about these cases seriously. The view is, however, very controversial (Levy 2015; Madva 2016).

Mandelbaum cites experimental evidence that implicit attitudes interact with other representational states in ways that indicate sensitivity to their semantic content. Consider, for example, the phenomenon of celebrity contagion (Newman, Diesendruck and Bloom 2011). People have positive implicit attitudes toward celebrities, and these attitudes seem to spread to their possessions. Thus they are willing to pay more for a sweater that they are told has been worn by an admired celebrity than an identical sweater that has never been worn. We can explain celebrity contagion associatively: the magic of celebrity rubs off on things they touch. However, it is harder to explain associatively why the willingness to pay more dissipates if they are told that the sweater has been washed after the celebrity wore it. After all, our implicit attitudes toward laundry are positive, not negative. There is nothing in the associative account to explain why washing reduces the perceived value of the sweater. As Mandelbaum says, that seems to indicate a kind of unconscious inference.

However, evidence that implicit attitudes are not associations, because they are responsive to the semantic content of other attitudes, may not be sufficient to show that they are beliefs. The evidence Mandelbaum cites is to the point, because beliefs are indeed centrally characterized by the way in which they respond to evidence and in which they interact with other attitudes. Beliefs are inferentially promiscuous, as Stich (1978) puts it. Critics have argued that implicit attitudes fall short of being beliefs (even if they are not *mere* associations) because they lack *broad* and *systematic* evidence responsiveness (Levy 2015; Madva 2016). Implicit attitudes seem to exhibit some inferential sensitivity, but may be surprisingly unresponsive to evidence that the person, rightly, takes be conclusive reason for updating their explicit attitudes (Gregg, Seibt and Banaji 2006). Conversely, they may respond to evidence that the person, rightly, rejects as a reason for belief update (Han, Olson and Fazio 2006). They may also be *completely* insensitive to negations (Wegner 1984; Deutsch, Gawronski and Strack 2006; Hasson and Glucksberg 2006). These insensitivities may indicate that implicit attitudes belong to a different class of representations altogether.

There is, however, an explanation of these insensitivities available compatible with implicit attitudes' being beliefs. No belief exhibits perfect inferential sensitivity, and there are plenty of cases where beliefs fail to update in response to evidence even when they ought to. Thus, in order to distinguish beliefs from other attitudes without removing *all* beliefs that fail to update from their analysis, many philosophers claim merely that an attitude must be able to update in response to evidence, or able to feature in inferences to count as a belief (Levy 2014). The question then becomes *how much* inferential and evidential sensitivity an attitude must demonstrate to count as a belief, and how "respectable" those transitions must tend to be.⁵

The problem, of course, is that different kinds of attitudes, variously wound up with our emotions, desires, self-conceptions, and self-expectations, will be more or less sensitive to evidence or inference, and there is reason to think that attitudes measured by tests like the IAT – implicit attitudes – are particularly susceptible to these potentially irrationalizing forces. Many of the tests conducted by social psychologists are designed to measure subjects' attitudes on topics that make many people deeply uncomfortable. One's attitudes about, for instance, race, gender, religion, sex, age, or wealth are attitudes about topics that are highly socially salient, and these are topics that can raise aversive feelings and a motivation to protect one's self-presentation (see Nier (2005) for an attempt to avoid self-presentational concerns). When tests like the IAT reveal that individuals have implicitly biased attitudes that conflict with one's explicit attitudes, this has the potential to threaten one's self-conception and cause self-protective explicit responding. Under conditions like these, it is hardly surprising that our attitudes may fail to update appropriately in response to reasons that the agent finds (explicitly) convincing, or will fail to integrate properly with her other attitudes.⁶

Another question we might raise has to do with the use of terms like "evidence." While social psychologists and philosophers often refer to an attitude's "sensitivity to update in response to evidence," little discussion of what "evidence" means has been pursued in these debates. Many psychologists seem to be using the term "evidence" in an *access*, rather than in a *possession* sense, where the ability to cite the reasons why one is updating one's beliefs indicates access to one's evidence. If it is possible to possess evidence (and for that evidence to be playing a causal role in one's cognition) without having conscious access to it, then we might wonder whether the causal reasons why one has an implicit attitude in the first place (typically explained as due to repeated exposure) ought to also count as the agent possessing evidence for their implicit attitude, despite the agent lacking conscious access to this evidence, and thus lacking the ability to report it as her evidence. If implicit attitudes are attitudes supported by repeated exposure to information, then it is also unsurprising that a single instance of contrary evidence or a presentation of reasons not to hold the attitude would fail to extinguish or update these attitudes immediately (while something like counterconditioning might be more successful) (see Lyons (2016) for a related discussion of unconscious evidence). Related issues about the strength of one's evidence, and the role evidence plays in updating our attitudes "appropriately," deserve further investigation.

The current state of play in the literature therefore leaves us with two questions. The first concerns the nature of implicit attitudes. Are they beliefs, associations, or *sui generis* states? There are a range of views in the literature, backed by evidence. It is probably fair to say, however, that the current state of the empirical literature doesn't allow for a definitive ruling on the question. The second question concerns the reliability of our intuitions about moral responsibility. If our intuitions are sensitive to the states recognized by folk psychology, as we've suggested, then they may be off-track with regard to these cases. That would threaten the use of what is still the standard methodology in debates concerning moral responsibility for the actions caused by implicit attitudes. Conceptual *and* empirical work is needed to return confident answers on both these (interlinked) questions.

3 Implicit Bias and Libertarianism

The big divide in theories of free will is, of course, between compatibilists and incompatibilists. Compatibilists hold that causal determinism is compatible with agential freedom, whereas incompatibilists deny this (roughly, a world is causally determined if every event

that occurs in it is caused by states of affairs or events that predate it, and there is no possible world with the same laws of nature and the same past as the determined world in which that event failed to occur). Compatibilism and incompatibilism are metaphysical theses: a compatibilist asserts that it is metaphysically possible that an action can be causally determined and free, while an incompatibilist denies this. In virtue of being a compatibilist, a philosopher is not committed to holding that there are agents who are free. Such a philosopher might think that in the actual world we lack freedom, for reasons other than determinism (e.g. Levy 2011). Compatibilists are, however, almost always concerned with how our agency is actually realized. So are libertarians (incompatibilists who hold that the actual world is not causally determined and agents in it are free).

To our knowledge, no one in the debate over implicit bias and agency has approached these issues from the perspective of debates over the compatibility thesis. Rather, they have assumed one or other compatibilist account of freedom. Doing so is justifiable, since the conditions set down by compatibilists account are widely held to be necessary (though not of course sufficient) by libertarians. However, there is room for a libertarianism that exploits implicit bias as an element of its account.

Consider Robert Kane's highly influential event-causal libertarianism (Kane 1996, 1999). Event-causal libertarians hold that agents act freely when their actions are indeterministically caused by appropriate agent-involving events (and other conditions, typically shared with compatibilists, are satisfied). Event-causal libertarianism, in common with other libertarian accounts of freedom, faces the luck objection. The luck objection is the following problem: if an agent's actions are undetermined, such that it is metaphysically possible that they ϕ or ψ in the prevailing circumstances, and nothing about them (their desires, wants, values, thoughts, and so on) settles which possibility becomes actual, then it is a matter of luck whether they ϕ or ψ . But it is difficult to see how someone may be free in virtue of something that is merely lucky for them. One of Kane's central innovations is his proposal for a solution to the luck objection. Kane holds that only in rare circumstances – only when she is very significantly torn as to how to act – is it metaphysically open to the agent whether to ϕ or to ψ . Agents are sufficiently torn between options when either is compatible with her values and other mental states. In these cases, she may try to perform *both* (mutually excluding) actions. Kane hypothesizes that when this occurs, the quantum level indeterminacy that characterizes neural processes is magnified by the resulting conflict, and it is this magnification of indeterminacy that makes either action metaphysically possible. Kane accepts that nothing about the agent settles how she acts (she is trying her hardest to choose both options, after all). However, he denies that how she acts is susceptible to a responsibility-undermining luck objection. However she acts, she was trying to act that way; she is therefore responsible for the action.

Most of the criticism of Kane's view has focused on the adequacy of the account as a response to the luck objection (Mele 2006; Levy 2011). However, its empirical plausibility may also be questioned. The speculation that conflict magnifies quantum level indeterminacy is no more than that. This is a thin reed to hang freedom from. Much the same can be said for David Hodgson's (2012) libertarianism. Hodgson's theory turns on the alleged capacity of conscious processes to grasp information as unified gestalts and thereby contribute to undetermined and non-formalizable decision-making. This power to make decisions that cannot be described algorithmically or subsumed under rules is the power to act freely, he maintained. But there is little evidence that the kind of conscious processes Hodgson appeals to have these powers. Event-causal libertarians might do better to appeal to implicit processes to build their theories.

Pace Hodgson, effortful, type-two, processing seems to be preferentially the domain of rule-governed cognition. Conversely, manipulations which increase the extent to which information processing is driven by implicit processes *decrease* the extent to which reasoning is algorithmic (De Neys 2006). These processes are generally blind to the structure of formal systems, but it is this structure that is essential to rule-based reasoning. These processes appear to respond to semantics but not syntax: even two-word phrases cannot be used as primes as a unit (Baumeister and Masicampo 2010). In contrast, rule-based reasoning often reflects “mindware” which is effortfully based and requires conscious attention for its application.⁷

These facts might provide elements for the construction of a new event-causal libertarianism. This libertarianism would be highly speculative, but is built on claims for which there is more supportive evidence than the accounts offered by Kane and by Hodgson. There is genuine (albeit patchy) evidence that implicit processes are stochastic, if not actually indeterministic. It is a relatively small inferential leap to postulating that they are genuinely indeterministic. They might provide the locus of libertarian freedom. Note, too, that such a libertarianism can borrow from Kane much of its response to the luck objection. Above we noted that implicit attitudes can have a decisive influence on how we act, bringing it about (say) that we choose candidate A rather than candidate B, only in circumstances in which small influences can serve to break (near) ties. To say that an agent faces a tie is to say that either option is such that it is compatible with her values, her beliefs, and so on. Her agency is, at minimum, a structuring cause of her options in these circumstances. In such circumstances, the luck objection loses some of its force since we may plausibly claim that either option is such that she that may identify with it.

4 Implicit Bias and Compatibilism

We noted above that most accounts of free will accept that the kinds of conditions identified by compatibilists are genuinely necessary, if not sufficient for freedom. They all therefore have a stake in the development of a plausible compatibilism. In this section, therefore, we turn to an assessment of the extent to which actions which are significantly influenced by implicit attitudes satisfy the conditions laid down by the most powerful compatibilist theories. We will focus, first, on the control condition, which has traditionally been the condition on which compatibilists have focused. Subsequently, we will turn to the attributability condition, which has recently been worked up into a powerful and plausible compatibilism to rival those centered around control.

Following Fischer and Ravizza (1998), we can understand control as consisting in patterned sensitivity to reasons. That is, an agent or a mechanism exercises control over a state of affairs if she or it is causally connected to that state of affairs and has the capacity to alter it in systematic response to reasons. Again following Fischer and Ravizza, we can cash this out in terms of counterfactuals: causal connection to the state of affairs is manifested in counterfactual circumstances in which the state of affairs responds to changes in the agent or mechanism; responsiveness to reasons is manifested in counterfactual circumstances in which changes in the state of affairs are explained by the agent’s or mechanism’s grasp of reasons for altering it. Finally, this responsiveness to reasons is systematic if the relevant counterfactuals are distributed systematically.

It will be helpful to provide an example of patterned sensitivity to reasons. A skilled driver who controls her car exhibits such sensitivity. The movements of the car are causally connected to her movements in obvious ways. Her responsiveness to reason is systematic in that

she is capable of ensuring that the car responds appropriately to fine-grained differences in reasons. As road conditions change, she alters her speed; she is capable of simultaneously adjusting speed and direction to pass other cars, perhaps at high speed, and so on. Obviously control understood as patterned sensitivity to reasons comes in degrees. An ordinary driver has a higher degree of control over his car than a novice driver, who responds only to roughly individuated reasons, but less control than a professional race car driver who responds appropriately to reasons to alter her momentum that the ordinary driver may fail even to perceive.

It is important to notice that control does not require conscious awareness of reasons or of its exercise. Conscious perception lags behind expert responsiveness to reasons in fast-moving environments, such as on the sports field. Even when we are monitoring our movements, we may fail to be aware of the corrections we make in response to reasons (Fournieret and Jeanerod 1998). Consciousness is a limited resource; we save it for the big picture and largely delegate the implementation details to nonconscious systems. In fact, conscious attention to these details may actually degrade performance (Beilock and Carr 2004). While the explicit/implicit distinction may cross-cut the conscious/unconscious distinction, the fact that processes that lack many of the properties we associate with consciousness help realize control should make us wary of claims that implicit processes may not (also?) play this role.

If implicit attitudes are beliefs – or play the functional role of beliefs – then it is very plausible that they realize what Haji (2012) calls “responsibility-level” control. Beliefs are systematically and directly responsive to reasons: it is this responsiveness to reasons that underlies and explains their inferential promiscuity. Beliefs interact with other states in the kinds of normatively respectable ways that *constitute* such reasons-responsiveness. However, as we saw above it is far from settled whether implicit attitudes are beliefs. If they are not, it matters how far they depart from being beliefs, insofar as their role in behavior is concerned. Control comes in degrees; Fischer and Ravizza maintain that responsibility-level control requires *moderate*, not *strong*, reasons-responsiveness. Furthermore, since implicit attitudes do seem to be subject to at least some degree of indirect reasons-responsive control (for instance, say that the agent learns about her implicit attitudes and then takes steps to change or extinguish them) (see Holroyd and Kelly 2016 for discussion), there is yet further reason to suspect that implicit attitudes may satisfy Fischer and Ravizza’s control condition. It is therefore possible that the control implicit attitudes help realize is sufficient for agential freedom, even if they are not beliefs.

We also saw that even if implicit attitudes *are* beliefs, they may fail to play the full functional role associated with beliefs (that of updating in response to evidence and featuring in inferences). Whether agents are responsible for the actions they cause would depend in part on *why* they don’t play that role.

Indirect tests like the IAT are often understood to be measuring automatic and uncontrollable reactions, which in turn are taken to indicate the existence of an implicit attitude. It’s tempting to take this as a straightforward indication that we’re not responsible for our implicit attitudes, but it’s important to distinguish different ways that we might control our attitudes and the actions that stem from those attitudes, and from there to determine what really matters for responsibility-level control. When discussing control and implicit attitudes, we might be referring to a lack of control over (i) the formation of the attitude (ii) the attitude itself, or (iii) the expression of that attitude in behavior. On most theories of moral responsibility, of central concern is the agent’s control (or lack thereof) over the behavior that the attitude issues in. This dovetails nicely with the empirical literature on implicit attitudes, since when psychologists say that implicit attitudes are difficult or impossible to control, they are typically claiming is that it is difficult or impossible to control (and typically, to inhibit) the

expression of these attitudes in behavior (take, for example, Fazio et al.'s view that indirect tests provide a "bona fide pipeline" to a person's true attitudes (1995)).

But is it true that implicit attitudes are uncontrollable in this way, as Fazio et al. (1995) claim? Contrary to claims like these, a number of studies have in fact shown that individuals are quite adept at controlling the manifestations of their attitudes in their behavior, even on indirect tests.

For instance, given instructions, participants are generally quite good at "faking" their results on some indirect tests, like the IAT, by manipulating how long it takes them to respond to the presented stimulus (Lai et al. 2014). In addition, a study by (Conrey et al. 2005) has shown that study participants can generally suppress biased reactions, if given the goal to do so (Gawronski, LeBel and Peters 2007, p. 186).⁸ Moreover, a study on shooter bias by Correll et al. (2007) showed that providing individuals with shooting training improves their capacity to control their automatic shooting responses in response to an overarching goal (a goal like: shoot people with guns; don't shoot people without guns).

What these studies appear to show, collectively, is that with training, motivation, and a higher-order goal to do so, agents might be able to control their behaviors on indirect tests. Analogously, should they wish to do so, and with attention, individuals might be able to suppress the expression of their biased attitudes in their behavior in everyday situations. Results like these suggest that social training of individuals, in addition to motivating to individuals to be less biased (give people some information in addition to a goal, something like: "Many of us harbour biases we aren't fully aware of. Try to be egalitarian! Check your biases!"), may result in individuals being able to successfully control their own implicitly biased behaviors, even if they cannot prevent themselves from forming the biased attitudes in the first place.

Of course, the opportunity to control one's responses on indirect tests will often depend on whether we know (in some sense) that we are engaging in or likely to engage in some behavior on that test, and what the significance of that behavior is. Thus, if we're responsible for the actions caused by our implicit attitudes, then this probably depends on our ability to know what attitudes we have, and how they're likely to influence our behavior in that situation. If this is right, then a good deal of self-knowledge is required for responsibility-level control over our actions caused by implicit attitudes. The degree and reliability of self-knowledge required for responsibility in these cases may be too high for individuals to count as responsible for these behaviors.

There are reasons to suspect that implicit attitudes reduce control below the threshold required for freedom, at least in those contexts in which we are most interested in implicit attitudes. We are most interested, we think, in the following set of cases: those in which there is a divergence between how the agent would have acted were her explicit attitudes to have controlled her behavior and how she actually acted (see Zheng 2016 for discussion). We are less interested in cases in which the implicit attitude brings it about that the agent acted as she would have were she fully informed and had the opportunity for reflection (in part because there is little unfairness in, say, accusing a wholehearted racist of engaging in racist action). In these contexts, however, it is the reduction in reasons-responsiveness that *explains* the character of the actual action. Had the agent been able to respond to the prevailing reasons, she wouldn't have acted in that manner. That seems a reason for thinking that the agent or the mechanism manifests a degree of control insufficient for freedom.⁹ Suppose, for instance, an agent is motivated to control her implicit attitudes and always makes a good faith effort to do so, but fails on some occasion for reasons for which she not responsible (she is under heavy load, for instance). As a consequence, she performs an action that she would not have had she succeeded in controlling her implicit attitudes. Her performance of the action seems to stem from a loss of control

over her behavior. Of course, her action *also* manifests a great deal of control (she responds to reasons in acting as she does), but the fact that it stems from a loss of control seems to entail a reduction in responsibility. Whether this reduction provides her with an excuse or instead mitigates her blame to some degree will depend on the threshold at which responsibility-level control is exhibited.

We turn now from control-based compatibilism to attributability-based accounts. On this family of accounts, agents are responsible for their actions if these actions are caused (non-deviantly) by attitudes that are deeply attributable to the agent. Proponents of these accounts have been explicit in maintaining that one of their attractions is that they enable us to hold agents responsible for actions caused by reasons of which they are not conscious. Consider Huck Finn (Arpaly 2003). Huck is praiseworthy for helping Jim, the escaped slave, avoid recapture, even though he consciously judges that his action is wrong. His actions are nevertheless caused by states that are deeply attributable to him. Angela Smith (2004, 2012; see Zheng 2016) has suggested that an account of this kind may encompass implicit attitudes. The implicit racism of a person might reflect – or constitute – her unconscious assessment of the worth of minorities, as much as her explicit nonracist attitudes may reflect her conscious evaluations.

Assessing the plausibility of these accounts requires us to assess the extent to which our implicit attitudes are deeply attributable to us. That, in turn, raises difficult questions concerning the deep self. On attributability views, we are responsible for our actions insofar as they reflect our grasp of reasons. On one set of views, the deep self can be identified, at least for these purposes, with the self for whom considerations count as reasons (see Brownstein 2016; Holroyd and Kelly 2016 and Levy 2017b for discussion and some dissent). A consideration counts as a reason for a person just in case it meshes with the set of states that constitutes her deep self. We might further flesh this out following Ismael (2015), identifying the deep self with the agent's *deliberative standpoint*. The deliberative standpoint is the standpoint from which the agent consistently identifies and responds to reasons across time. Something counts as a reason for her only if she takes it to be one from this standpoint. Here, taking something to be a reason is a functional notion: a consideration is taken as a reason if it plays the right role: disposing the agent to action or changing her credences (note that a consideration need not be endorsed to play these roles. An agent might deny that a certain fact is a genuine reason while it nevertheless causes her to update her take on things, for example).

The evidence of the patchiness with which implicit attitudes play this functional role could be interpreted as showing that they don't – or only patchily – constitute agents' deliberative standpoint. One issue that deserves attention is whether such standpoints must be unique and consistent. Insofar as they are supposed to be identical to, or play an important role in constituting, the agent, there is pressure to think of them in these terms. It is unintuitive to attribute conflicting attitudes to one and the same agent, let alone to say she is identical to or constituted by such attitudes. But we have already seen that there may be reasons to regard our intuitions on these topics as unreliable.

5 Conclusion

Philosophical work on implicit bias is highly demanding, insofar as it requires both empirical sophistication and a deep knowledge of relevant philosophical debates. A number of philosophers have risen to the challenge it presents and there are already a number of differing, sometimes complementary sometimes conflicting perspectives on the nature of implicit

bias and its normative import. We have offered an opinionated introduction to those parts of this debate that bear most directly on moral responsibility and free will. Even on these topics, there is a great deal more to be said (there is, for instance, fascinating work on the effects of holding people responsible for actions caused by implicit attitudes; see Saul 2013; Zheng 2016). Moreover, this is a rapidly shifting terrain, since it is one in which philosophical debates must respond to developments in a contested science.

We have not defended a view here. We have different takes on the overall drift of the literature and agree that however our disagreements play out, much more data, and more philosophical work, is needed to settle these debates. We still cannot be confident about the nature of implicit attitudes or the extent to which they are identified with conscious states. Without definite answers to these questions, we cannot be confident whether agents satisfy the conditions laid down by standard conceptions of moral responsibility.

Notes

- 1 There is room for skepticism that even these debates can be entirely insulated from the cognitive sciences. We build theories in metaphysics in ways that are constrained and guided by our intuitions. But some intuitions may be unreliable. For example, our intuition that the outcomes of actions are relevant to the culpability of the agent, independently of her intentions, may be the product of mechanisms designed to give us information about agents, such that if we give it any weight when we have independent information about her intentions, we double count the intention and inappropriately increase our assessment of her culpability (Levy 2016). If evidence from cognitive science gives us reason to regard some of our intuitions as about the compatibility of freedom and determinism (say) as unreliable, we would have reason to take it into account in the relevant debates.
- 2 Note that the distinction between attitudes available to report and those that are not cross-cuts the endorsed/unendorsed attitude distinction. People may report their “gut feelings” without endorsing them. On some views, implicit attitudes are realized by gut feelings (perhaps among other states). Here we take no stand on whether this is defensible.
- 3 It’s important to note that implicit attitudes may be (more or less) *consistent* with one’s explicit attitudes. In these cases, individuals will not exhibit attitude dissociation and the issue of endorsement doesn’t clearly arise (cf. Greenwald and Banaji: “[T]he occurrence of dissociation is not a necessary condition for identifying an attitude as implicit” (1995, p. 9)).
- 4 Note that if intuitions are generated by domain-specific mechanisms, we should expect them to be recalcitrant to correction. Domain-specific mechanisms are, by definition, insensitive to considerations available to the agent which fall outside the domain to which they are attuned. Hence, the intuitions will persist, even when we judge them to be inappropriate. Compare visual illusions, which are generated by domain-specific perceptual mechanisms. The knowledge that the lines in the Müller-Lyer illusion are the same length doesn’t affect these mechanisms, and therefore we continue to have the perceptual seeming that they are different lengths.
- 5 Others claim that the ability to update in response to evidence is still too strong of a condition on belief, and that implicit attitudes are not beliefs because of their insensitivity to (for instance) negation (see Madva (2016)).
- 6 Another possible explanation of these cases is that agents lack self-knowledge, and their explicit self-expressed attitudes are more like self-*projections* than self-*reports*.
- 7 Whether Hodgson was wrong to associate this kind of processing with consciousness remains an open question: it may be that implicit processes are partially conscious. While most early theorists assumed that agents lacked awareness of their implicit attitudes, there is evidence that at least some agents are aware of the content of some of these attitudes. For instance, Nier (2005) told experimental participants that attempts to misrepresent their racial attitudes would be detected by experimenters; these participants reported explicit attitudes that were much better correlated with

their implicit attitudes than is usually seen in experiments like this. Similarly, Ranganath, Smith and Nosek (2008) asked subjects to rate their “gut reactions” and their “actual feelings” toward gay people. People reported “gut reactions” that were more negative than their “actual feelings”; moreover, their gut reactions correlated well with their implicit attitudes. The implicit/explicit distinction cross-cuts the unconscious/conscious distinction (to the extent that’s true, it is a mistake to appeal to conscious processes as Hodgson does).

- 8 Gawronski, LeBel, and Peters (2007) cite a number of other studies that they claim also show this result, including Conrey et al. (2005); Klauer and Teige-Mocigemba (2007); Lowery, Hardin, and Sinclair (2001); and Payne (2001).
- 9 Note that Fischer and Ravizza would not regard such a local failure of reasons-responsiveness to be exculpating, so long as the *overall pattern* of reasons-responsiveness is systematic (enough). This seems to us to be mistaken, given that it is the failure of reasons-responsiveness that explains the character of the action. See Levy (2017a) for discussion.

Bibliography

- Arpaly, N. (2003). *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.
- Baumeister, R.F. and Masicampo, E.J. (2010). Conscious thought is for facilitating social and cultural interactions: how simulations serve the animal-culture interface. *Psychological Review* 117: 945–971.
- Beilock, S.L. and Carr, T.H. (2004). From novice to expert performance: attention, memory, and the control of complex sensorimotor skills. In: *Skill Acquisition in Sport: Research, Theory and Practice* (ed. A.M. Williams, N.J. Hodges, M.A. Scott and M.L.J. Court), 309–328. London: Routledge.
- Brownstein, M. (2016). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology* 7: 765–786.
- Conrey, F.R., Sherman, J.W., Gawronski, B., Hugenberg, K., and Groom, C.J. (2005). Separating multiple processes in implicit social cognition: the quad model of implicit task performance. *Journal of Personality and Social Psychology* 89 (4): 469–487.
- Correll, J., Park, B., Judd, C.M., Wittenbrink, B., Sadler, M.S., and Keese, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology* 92 (6): 1006–1023.
- De Neys, W. (2006). Dual processing in reasoning: two systems but one reasoner. *Psychological Science* 17: 428–433.
- Deutsch, R., Gawronski, B., and Strack, F. (2006). At the boundaries of automaticity: negation as reflective operation. *Journal of Personality and Social Psychology* 91: 385–405.
- Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.
- Evans, J.S.B.T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59: 255–278.
- Faucher, L. (2016). Revisionism and moral responsibility for implicit attitudes. In: *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics* (ed. M. Brownstein and J. Saul), 115–144. Oxford: Oxford University Press.
- Fazio, R.H., Jackson, J.R., Dunton, B.C., and Williams, C.J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? *Journal of Personality and Social Psychology* 69 (6): 1013–1027.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control: An Essay on Moral Responsibility*. Cambridge: Cambridge University Press.
- Fourneret, P. and Jeannerod, M. (1998). Limited conscious monitoring of motor performance in normal subjects. *Neuropsychologia* 36: 1133–1140.
- Gawronski, B., LeBel, E.P., and Peters, K.R. (2007). What do implicit measures tell us? scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science* 2 (2): 181–193.
- Greenwald, A.G. and Banaji, M.R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review* 102 (1): 4–27.

- Greenwald, A.G., Banaji, M.R., and Nosek, B.A. (2015). Statistically small effects of the implicit association test can have societally large effects. *Journal of Personality and Social Psychology* 108: 553–561.
- Greenwald, A.G., McGhee, D.E., and Schwartz, J.K.L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology* 74: 1464–1480.
- Greenwald, A.G., Poehlman, T.A., Uhlmann, E., and Banaji, M.R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97: 17–41.
- Gregg, A.P., Seibt, B., and Banaji, M.R. (2006). Easier done than undone: asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology* 90: 1–20.
- Haji, I. (2012). *Reason's Debt to Freedom*. Oxford: Oxford University Press.
- Han, H.A., Olson, M.A., and Fazio, R.H. (2006). The influence of experimentally-created extrapersonal associations on the implicit association test. *Journal of Experimental Social Psychology* 42: 259–272.
- Hasson, U. and Glucksberg, S. (2006). Does negation entail affirmation? the case of negated metaphors. *Journal of Pragmatics* 38: 1015–1032.
- Hodgson, D. (2012). *Rationality + Consciousness = Free Will*. Oxford: Oxford University Press.
- Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy* 43: 274–306.
- Holroyd, J. and Kelly, D. (2016). Implicit bias, character, and control. In: *From Personality to Virtue* (ed. J. Webber and A. Masala), 106–133. Oxford: Oxford University Press.
- Howell, J.L., Collisson, B., Crysel, L., Garrido, C.O., Newell, S.M., Cottrell, C.A., Smith, C.T., and Shepherd, J.A. (2013). Managing the threat of impending implicit attitude feedback. *Social Psychological and Personality Science* 4 (6): 714–720.
- Ismael, J.T. (2015). On being someone. In: *Surrounding Free Will: Philosophy, Psychology, Neuroscience* (ed. A.R. Mele), 274–297. Oxford: Oxford University Press.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.
- Kane, R. (1999). Responsibility, luck, and chance: reflections on free will and indeterminism. *Journal of Philosophy* 96: 217–240.
- Klauer, K.C. and Teige-Mocigemba, S. (2007). Controllability and resource dependence in automatic evaluation. *Journal of Experimental Social Psychology* 43 (4): 648–655.
- Lai, C.K., Marini, M., Lehr, S.A., Cerruti, C., Shin, J.E.L., Joy-Gaba, J.A., Ho, A.K. et al. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General* 143 (4): 1765–1785.
- Levy, N. (2011). *Hard Luck*. Oxford: Oxford University Press.
- Levy, N. (2014). *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.
- Levy, N. (2015). Neither fish nor fowl: implicit attitudes as patchy endorsements. *Noûs* 49: 800–823.
- Levy, N. (2016). Dissolving the puzzle of resultant moral luck. *Review of Philosophy and Psychology* 7: 127–139.
- Levy, N. (2017a). Implicit bias and moral responsibility: probing the data. *Philosophy and Phenomenological Research* 94: 3–26.
- Levy, N. (2017b). Am I racist? Implicit bias and the ascription of racism. *Philosophical Quarterly* 67: 534–551.
- Lowery, B.S., Hardin, C.D., and Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology* 81 (5): 842–855.
- Lundberg, K.B. and Payne, B.K. (2014) Decisions among the undecided: implicit attitudes predict future voting behavior of undecided voters. *PLoS One* 9 (1): e85680. doi: 10.1371/journal.pone.0085680.
- Lyons, J. (2016). Unconscious evidence. *Philosophical Issues* 26: 243–262.
- Madva, A. (2016). Why implicit attitudes are (probably) not beliefs. *Synthese* 193: 2659–2684.
- Mandelbaum, E. (2016). Attitude, inference, association: on the propositional structure of implicit bias. *Noûs* 50: 629–658.
- Mele, A. (2006). *Free Will and Luck*. New York: Oxford University Press.
- Newman, G., Diesendruck, G., and Bloom, P. (2011). Celebrity contagion and the value of objects. *The Journal of Consumer Research* 38: 215–228.

- Nier, J.A. (2005). How dissociated are implicit and explicit racial attitudes?: a bogus pipeline approach. *Group Processes & Intergroup Relations* 8: 39–52.
- Nosek, B.A., Greenwald, A.G., and Banaji, M.R. (2007). The implicit association test at age 7: a methodological and conceptual review. In: *Automatic Processes in Social Thinking and Behavior* (ed. J.A. Bargh), 265–292. New York: Psychology Press.
- Oswald, F.L., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P.E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology* 105: 171–192.
- Payne, B.K. (2001). Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology* 81 (2): 181–192.
- Payne, B.K., Krosnick, J.A., Pasek, J., Leikes, Y., Akhtar, O., and Tompson, T. (2010). Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology* 46: 367–374.
- Pereboom, D. (2013). Free will skepticism, blame and obligation. In: *Blame: Its Nature and Norms* (ed. D.J. Coates and N. Tognazzini), 189–206. Oxford: Oxford University Press.
- Ranganath, K., Smith, C., and Nosek, B. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology* 44: 386–396.
- Saul, J. (2013). Unconscious influences and women in philosophy. In: *Women in Philosophy: What Needs to Change?* (ed. F. Jenkins and K. Hutchison), 39–60. Oxford: Oxford University Press.
- Schwitzgebel, E. (2001). In-between believing. *Philosophical Quarterly* 51: 76–82.
- Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly* 91: 531–553.
- Smith, A. (2004). Conflicting attitudes, moral agency, and conceptions of the self. *Philosophical Topics* 32: 331–352.
- Smith, A. (2012). Attributability, answerability, and accountability: in defense of a unified account. *Ethics* 122: 575–589.
- Stich, S. (1978). Beliefs and subdoxastic states. *Philosophy of Science* 45: 499–518.
- Storage, D., Charlesworth, T.E.S., Banaji, M.R., and Cimpian, A. (2020). Adults and children implicitly associate brilliance with men more than women. *Journal of Experimental Social Psychology*. doi: 10.1016/j.jesp.2020.104020.
- Uhlmann, E.L. and Cohen, G.L. (2005). Constructed criteria: redefining merit to justify discrimination. *Psychological Science* 16: 474–480.
- Washington, N. and Kelly, D. (2016). Who's responsible for this? implicit bias and the knowledge condition. In: *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics* (ed. M. Brownstein and J. Saul), 11–36. Oxford: Oxford University Press.
- Wegner, D. (1984). Innuendo and damage to reputation. *Advances in Consumer Research* 11: 694–696.
- Zheng, R. (2016). Attributability, accountability and implicit attitudes. In: *Implicit Bias and Philosophy: Volume 2, Moral Responsibility, Structural Injustice, and Ethics* (ed. M. Brownstein and J. Saul), 62–89. Oxford: Oxford University Press.

The Role of Consciousness in Free Action

PHILIP WOODWARD

1 Introduction

It is easy to prompt the intuition that free will and consciousness have something importantly to do with one another. Consider cases such as the following:

Case 1: Your spouse accuses you of having broken your commitment to the family diet by stealing cake from the refrigerator in the middle of the night, and he is in possession of damning videographic evidence. In your defense, you point to your long history of sleep-walking, and you suggest that you make an appointment with a sleep doctor.

Case 2: On a weekend, you get behind the wheel of your car and head to the grocery store. As you drive, your mind drifts, until you suddenly realize that you are *not* on your way to the store after all, but have instead followed the route you take to work each weekday.

In neither of these cases does it seem that your action (eating a piece of cake, driving to work) is the product of your free will, and this fact can be explained, in both cases, in terms of your lack of conscious awareness of what you were doing. But it is not always so intuitively clear what type or degree of conscious awareness is required for an action to count as free, as exemplified in the following cases:

Case 3: During a game of basketball, you steal the ball and make a fast break down the court. As you set up to make a shot, you spot an open teammate out of the corner of your eye, and without thinking dish the ball to her so she can complete an unimpeded lay-up.

Case 4: As you watch an advertisement on television for a charity organization, you resolve to make a generous donation, and this in contrast to your typical stinginess. Later you learn that the advertisers flashed the word “GIVE” for a duration too brief for you to have consciously processed.

In Cases 3 and 4, your actions (passing a basketball, making a donation) are significantly shaped by non-conscious processes of which you were unaware at the time of acting. As a consequence, should we say that they were not done freely?

This question is difficult to answer in the absence of a general theory of the relationship between consciousness and free will. And no consensus currently exists regarding this relationship. The reason is that the cognitive function of consciousness, on the one hand, and the necessary conditions for freedom, on the other, are matters of considerable controversy. Regarding the cognitive function of consciousness, at least the following five options have contemporary adherents (roughly in descending order of how psychologically important consciousness is):

The Selfhood Role: A subject, properly speaking, is identical with a *conscious* subject; non-conscious states and processes play only a supporting role. (While held paradigmatically by robust realists about the self, this view is by no means restricted to them. It can also be held by those who identify the self with the stream of consciousness, or with a narrative constructed out of conscious experiences.)

The Self-awareness Role: Consciousness alone affords subjects with privileged access to their own mental states.¹

The Content-Grounding Role: All of a subject's contentful mental states are individuated via their relations to actual or potential conscious states; non-conscious states count as contentful (that is, being about something, representing something) – or even as *mental* – only derivatively.²

The Integration Role: Consciousness coordinates the functioning of otherwise self-contained cognitive modules, by integrating the information processed by these modules into a central, domain-general stream of information.³ (“Domain-general” contrasts with “domain-specific,” i.e., pertaining to one type of information – visual stimuli, for example).

The Grab-Bag Role: There is a set of loosely related cognitive processes such that consciousness is always or is typically involved in their implementation. Candidates include forming a self-concept; learning new, complex behaviors, building associations between temporally separated events; performing multi-step, goal-directed activities.⁴

Regarding the necessary conditions for freedom, the following four are the most commonly discussed:

The Control Condition: An action is free only if it is generated by the agent, rather than coerced or compelled by forces external to the agent.

The Reasons-Responsiveness Condition: An action is free only if it is done for a reason, or else such that if the agent had had a different or better reason, the agent would have acted according to that reason.⁵

The Identification Condition: An action is free only if it expresses the agent's “true self” (i.e. her established commitments/values) or is such that the agent can or would *identify* with the action.

The Open-Future Condition: An action is free only if temporally prior states of the agent and her environment, in conjunction with the laws of nature, do not determine that she performs it.⁶

(I mean these formulations only as initial, rough approximations. In what follows I will note when the differences between rival formulations become relevant.)

It's possible to combine items from each of these two lists to form a wide range of general theories about the relationship between consciousness and free will. In the next section, I discuss four (families of) options: the constitution view, the causal-dependence view, the counterfactual-dependence view, and the independence view. In the final section, I review some of the most influential recent findings in the empirical sciences that touch on the

relationship between consciousness and the will, and I discuss the potential philosophical significance of these findings.

Two caveats. First, some of the literature I draw on focuses on the relationship between consciousness and moral responsibility rather than freedom. Second, while my focus will be on *phenomenal* consciousness (mental states' subjective feel), some of the literature I draw on focuses on *access* consciousness (mental states' availability for participation in various online processes). The concept of moral responsibility is not the same concept as that of freedom, nor is the concept of phenomenal consciousness the same as the concept of access consciousness. Nevertheless, the extensions of the concepts in each pair overlap enough to serve my purposes.⁷

2 Conceptual and Metaphysical Considerations

In the present section, I discuss four ways of understanding the metaphysical or conceptual connection between consciousness and free action. The first three posit a modally strong connection of some sort, but differ in how direct that connection is. I will call these three views "the constitution view," the causal-dependence view," and the "counterfactual-dependence view." The fourth view, "the independence view," is the denial that any such modally strong connection exists.

1. *The Constitution View.* According to constitution view, necessarily, all free action is constituted by consciousness. That is, free actions metaphysically consist, at least in part, in phenomenally conscious episodes of a special sort. We can get the flavor of a view of this sort by putting together two observations about action – or rather, controversial gloss on two uncontroversial observations about action.

The first uncontroversial observation is that there is, at least sometimes, something that it is like to act – a phenomenology of action.⁸ More controversial is the claim that the phenomenology of action consists in an irreducibly "executive" type of phenomenology or as Ginet (1990) puts it, a distinctively "actish" phenomenal quality that cannot be reduced to other types of phenomenology, for example the conscious representation of oneself as acting, or to a consciously represented intention to act followed by feeling one's body move.⁹

The second uncontroversial observation is that, among the many correct descriptions of a token action (flip the switch, turn on the light, alert the intruder, etc.), some do a better job than others at capturing the agent's perspective on the action, and one description in particular will capture the most basic action that the agent is trying to perform.¹⁰ More controversial is the claim that basic actions can occur in cases of total action-failure: for example, a unknowing paralytic can *try* to raise her arm – and thus really *do* something – without producing the slightest muscular movement.¹¹ If we suppose that the concept of executive phenomenology and the concept of a basic action converge on the same psychological phenomenon, what emerges is the existence of what Brian O'Shaughnessy (1973, 2003) calls "tryings": conscious experiences that consist "in doing, intentionally and with just that purpose, whatever one takes to be needed if, the rest of the world suitably cooperating, one is to perform that action."¹²

Armed with this category of executive phenomenology, we can now understand what the constitution view comes to. According to the constitution view, free actions are constituted (at least in part) by conscious episodes of an irreducibly executive type. No action that does not partly consist in an agent's consciously trying to perform it counts as free.

Why think free action is so tightly connected to consciousness? There are, I think, two types of argument that capture the appeal of the constitution view. The first argument links

together the initial items on each of the two lists I presented above (proposed cognitive roles of consciousness, proposed necessary conditions on free action). Suppose that consciousness plays the *selfhood role*: a subject, properly speaking, is identical with a *conscious* subject. To this claim add the Control condition on freedom (an action is free only if it is generated by the agent, rather than coerced or compelled by forces external to the agent). Assuming that an *agent* simply is a self or subject who is (or is capable of) acting, then it follows from these two commitments that acts are free only if generated by a conscious subject. There is more than one way to understand what it is for a conscious subject, as such, to “generate” an act, but perhaps the simplest picture is one on which the act is a part of the subject’s stream of consciousness, which is what the constitution view maintains. (In a slogan: unless you’re directly conscious of what you’re doing, you’re not in control of it.) Call this the “Selfhood Argument” for the constitution view.¹³

The second argument links together the second items on each of the two lists above. Suppose that consciousness plays the *self-awareness role*: that is, consciousness alone affords subjects with privileged access to their own mental states. To this claim add the Reasons-Responsiveness condition on freedom (an action is free only if it is done for a reason, or else such that if the agent had had a different or better reason, the agent would have acted according to that reason.) These two commitments motivate the constitution view, given the additional claim that an agent (a) acts for some reason only if she has privileged access to the reason for which she acts, and (b) has such privileged access simply in virtue of performing the action. Though there may be more than one way for an agent to have privileged access to her reasons for acting “in virtue of” performing an act, perhaps the simplest way for this to occur is if executive phenomenology itself encodes the reason for which an action is consciously executed. That is, one is only rationally guiding one’s action if, simply in virtue of consciously acting, one is aware of one’s motivating reason for acting. (In a slogan: unless you’re directly conscious of why you’re doing what you’re doing, you’re not doing it for a reason.) Call this the “Self-Awareness Argument” for the constitution view.¹⁴

2. *The Causal-Dependence View*. According to this view, necessarily, all free action is causally dependent on consciousness. The causal-dependence view shares with the constitution view the commitment that consciousness plays a part in all free actions. But it is weaker than the constitution view, because it does not require that conscious events themselves count as actions. Rather, it is consistent with the view that free actions are cause/effect pairs, and only the causes need be conscious. For example, one version of this view would be that free actions consist in physical or mental changes caused by freely formed conscious intentions. If for whatever reason a conscious intention fails to effect any change, no action has been performed (either because utterly unsuccessful tryings are impossible, or because they are possible but are not in themselves actions).

The causal-dependence view can be motivated by slightly weaker versions of the Selfhood Argument and the Self-Awareness Argument. According to the weaker version of the Selfhood Argument, all that is required for a conscious subject, as such, to generate an action is for the triggering cause of that action to be part of her stream of consciousness (the action itself need not be). And according to the weaker version of the Self-Awareness Argument, all that is required for an agent to have privileged access to her reasons for acting is for the intention that governs her action to be part of her stream of consciousness (the action itself need not be).

The Open-Future condition provides a third motivation for the causal-dependence view. According to the Open-Future condition, recall, an action is free only if temporally prior states of the agent and her environment, in conjunction with the laws of nature, do not

determine that she performs it. Many philosophers have supposed that the Open-Future condition is in *prima facie* tension with the Control condition: if an agent's actions are not determined by temporally prior states of hers, then how is she in control of her actions? Libertarians have responded to this challenge in a variety of ways.¹⁵ One such response that appeals directly to consciousness is the so-called "wave-function collapse" theory of mental causation, developed by Henry Stapp (2005) and Hans Halvorsen (2011).¹⁶ Physical systems can inhabit "superpositions": they can lack determinate values for some of their quantities. But, plausibly, it's not possible for conscious states to exhibit that sort of indeterminacy. According to the wave-function collapse theory, physical systems come to inhabit determinate states when they are suitably causally related to determinate, conscious states. Thus, contemporary physics may leave room for consciousness to play a fundamental role in setting the parameters for the evolution of an otherwise indeterminate, open-ended system. Advocates of the wave-function collapse theory provide support for their view by arguing that it solves the thorny "measurement problem" in quantum mechanics. Nevertheless, it remains a minority view.

There are reasons to think that both the constitution view and the causal-dependence view are too restrictive. Suppose an agent works hard to overcome a compulsion to look at his cell phone while driving. His inclination is to reach for his phone when it vibrates, but he conscientiously develops an alternative habit – giving the steering wheel a strong squeeze, say. Eventually he reacts in this way even when his attention is otherwise occupied. Is his action unfree? Certainly his action meets the Identification condition: his behavior is the manifestation of his established commitments and values.¹⁷

Consider another example. Most drivers are not aware of signaling when turning; they simply intend to turn, and signal that they are about to turn as a matter of habit.¹⁸ At the same time, their signaling is not a haphazard or reflexive action, but is done for a reason, namely, in order to alert other drivers to an immanent directional change. Thus when drivers signal, they satisfy the Reasons-Responsiveness condition. (And, of course, if they fail to signal, they can be held legally and morally responsible.)

There are two ways for the advocate of the causal-dependence view to accommodate these cases. First, she could maintain that, while the Identification and Reasons-Responsiveness conditions are necessary conditions on freedom, they are not sufficient; conscious causation of action is required as well. After all, the purpose of cultivating habits is to establish a new default way of behaving. And in an important sense, default behaviors are not freely chosen.

A second way of responding is to grant that the cases just described can be instances of free actions, so long as conscious control has played a role in the cultivation of the relevant habits. That is, the habits in question were acquired by the agent only as he has engaged in repeated, consciously controlled action. So even if his current action (squeezing the wheel, signaling a turn) is not consciously caused, it is the causal descendent of similar actions that *were* consciously caused. To accommodate this response, we can distinguish between two versions of the Control condition. According to the first, the exercise of control has to be *direct* (i.e., proximal):

Direct Control Condition: An action is free only if it is directly generated by the agent.

According to the second, control can be exercised in an indirect way, via the direct cultivation of a habit or the forming of an intention to carry out a long-term project:

Indirect Control Condition: An action is free only if it is directly generated by the agent or is the manifestation of a habit or standing intention that was directly generated by the agent.

These two versions of the Control condition can be used to motivate different versions of the causal-dependence view. On the strong causal-dependence view, an action is free only if it is caused by a conscious intention on the part of the agent. On the weak causal-dependence view, an action is free only if it is caused by a conscious intention on the part of the agent, or it is the manifestation of a habit that has been cultivated as the causal consequence of conscious intentions on the part of the agent.¹⁹

3. *The Counterfactual-Dependence View.* According to this view, necessarily, all free action is counterfactually dependent on consciousness. Returning to the two cases of habitual action just discussed: there is another way to understand how these cases satisfy the Control condition, a way that doesn't appeal to the agent's history. Both the direct and indirect versions of the Control condition are *actualist* versions, but the Control condition can also be understood counterfactually:

Counterfactual Control Condition: An action is free only if the agent would have generated the action had automatic mechanisms been insufficient for its occurrence.

If the Control condition is understood in this way, then it generates an even weaker version of the Selfhood Argument. Supposing, again, that consciousness plays the selfhood role (such that only conscious action-generation counts as genuine control), it follows from the Counterfactual Control condition that an action is free only if it *could have been* consciously generated.

In the cases above, non-conscious wheel-squeezing and non-conscious signal-turning both satisfy this condition. That is, perhaps what makes these actions free (if they are) is not so much the agent's history but the agent's possessing, at the time of his acting, certain capacities that may or may not be operative. We can contrast such free-but-not-consciously-caused behaviors with compulsions (e.g., tics, behaviors following hypnotic suggestion, addictive behaviors) and with unwilled bodily movements (e.g., heart-beats, saccading of the eyes), neither of which can be controlled consciously.

Likewise, both variants of the Self-Awareness Argument we have considered so far have rested on an actualist construal of the Reasons-Responsiveness condition (just as the Selfhood Argument for the constitution view and the causal-dependence view rested on an actualist construal of the Control condition). But there is also a counterfactualist version of the Reasons-Responsiveness condition.

Counterfactual Reasons-Responsiveness Condition: An action is free only if, had the agent possessed a reason to act otherwise, the agent would have acted for that reason.²⁰

So understood, the Reasons-Responsiveness condition generates a weaker version of the Self-Awareness Argument. Suppose that consciousness plays the self-awareness role, and that an agent (a) acts for some reason only if she has privileged access to the reason for which she acts, and (b) has such privileged access simply in virtue of performing the action. It will follow from the Reasons-Responsiveness condition only that it must be *possible* for an agent to consciously entertain reasons for acting otherwise than she (freely) does, not that her actual reason for acting must be conscious at the time of her (freely) acting.

The Selfhood and Self-Awareness Arguments appeal to the first two proposed cognitive roles for consciousness (respectively). The counterfactual dependence view can be motivated independently via the third: namely, the content-grounding role. John Searle has famously argued that, necessarily, all mental states with intentional content are either actually or potentially conscious, a claim he calls "The Connection Principle." A "potentially conscious"

mental state just is a capacity to bring about a *conscious* intentional state.²¹ Or, as Declan Smithies (2012) puts the idea, all intentional states are either conscious states or they are individuated by their relations to conscious states. For example, on this view, what it is for a subject to believe that Cairo is the capital of Egypt is for a subject to be capable of consciously judging that Cairo is the capital of Egypt under certain circumstances. Though a belief isn't a conscious state, it is individuated via its relation to conscious states.

Searle and his ilk argue for the Connection Principle in several ways. First, they claim that on reflection, the idea of intentional content that that is not presented to a subject makes no sense.²² Second, they argue that all attempts to explain how the content of mental states is fixed, save for those that appeal to consciousness, leave content hopelessly indeterminate.²³ Third, they argue that all attempts to explain why some systems have intentional properties while other systems do not, save for those that appeal to consciousness, imply that unthinkable things such as tornadoes and toilets have intentional states.²⁴

If these arguments are successful, then whenever an agent is motivated by a reason to do something – i.e., if her mental states encode that reason in any way – then she must be capable of consciously entertaining the content of that reason. Thus, on any version of the Reasons-Responsiveness condition, it follows from the Connection Principle that an agent acts freely only when she is capable of consciously entertaining the content of her reason for acting. Now, whatever the merits of Searle's arguments, they have failed to move the majority of philosophers of mind, who maintain on the contrary that both folk and empirical psychology rightly invoke non-conscious intentional states that have no constitutive connection to consciousness.

Finally, as King and Carruthers (2012) point out, the Identification condition on its own motivates the counterfactual-dependence view. The Identification condition is substantive only if there is a way to distinguish between an agent's established commitments and values, on the one hand, and those attitudes of hers she does not genuinely identify with, on the other (compulsive or deeply selfish desires, say). Unless the agent *would consciously endorse* some attitude of hers, it doesn't seem right to treat that attitude as part of her established commitments and values.^{25,26}

The counterfactual-dependence view grants the metaphysical possibility of an agent whose conscious states are never causally responsible for his actions, but who nevertheless acts freely. We can draw an analogy with the relationship between a project-manager and the team of employees she oversees. We can imagine that the team is so effective that at no point does the manager ever exercise any top-down control, not even at the project's outset. (The employees are so deeply committed to the mission of the corporation that they can anticipate which projects need to be implemented before being told.) Nevertheless, the manager is poised to intervene *were* the team to get off-track. If we find it plausible to say in such a scenario that the manager controls the project and that the team is responsive to the manager's goals, then we are not likely to be troubled by the fanciful epiphenomenalist possibility that the counterfactual-dependence view permits.

4. *The Independence View.* The counterfactual-dependence view is consistent with life-long epiphenomenalism on the part of an agent's conscious states, so long as the agent has the *capacity* to be conscious in certain ways. But not all theorists about free will and consciousness are committed even to a counterfactual dependence relation between free action and consciousness. There are philosophical and scientific theories of agency, on the one hand, and of consciousness, on the other, that are not obviously mutually implicating in any way; any connections that actually obtain between them are contingent.

Consider, for example, biologically based theories of agency. According to these theories, what it is to be an agent is to be an "autonomous" or self-organizing system, i.e., a system that

maintains its homeostasis over and against, but also drawing upon, environmental contingencies. The more “flexibility” a system exhibits – that is, the less its behavior is dictated by its immediate environment – the more free it is. This distinctly biological notion of agentic control requires no contribution from the system’s stream of consciousness (if it even has one).²⁷

Certain computationally based theories of agentic control are likewise silent on the role of consciousness. According to one leading computational model, control consists in the successful deployment of “comparator” mechanisms. These mechanisms exploit a feedback loop linking (a) an initial action plan, (b) the predicted outcome of that action plan, and (c) the perceived outcome of that action plan.²⁸ On this model, an agent is in control of her action when these feedback loops successfully enable her to complete her action plan, updating it in real time if necessary – picking up one’s cup of coffee without spilling it, for example. In other words, an agent controls her action to the degree that her intentions are successfully implemented, and there is no requirement that the agent be aware either of her intentions or the means by which they are executed.

Now, there are two procedures implicated by model of control that might require the assistance of consciousness. First, consciousness might be required as part of the feedback loop. For example, it might be that conscious *attention* is what enables a subject to acquire the requisite perceptual feedback.²⁹ But it’s not clear why the model would require that such attention be *conscious*. For example, when a somnambulist successfully extracts a piece of cake from the refrigerator, she seems to be satisfying all of the model’s requirements for exercising control, yet her attention to the cake remains unconscious.³⁰

The second potential role for consciousness within the comparator model of control is in fixing those intentions against which successful control is measured. That is, perhaps controlled, reason-responsive, and/or true-self-expressive action only occurs when the agent’s intention was formed via conscious deliberation (which would take us back to the causal-dependence view), or at any rate such that the agent could have consciously entertained her intention (which would take us back to the counterfactual-dependence view).

As far I can tell, however, nothing in the computational model of control itself suggests these restrictions. Now, computationalists often do make a distinction between “person-level” representations and “sub-personal-level” representations. The former are those representations that play a role in explanations of the behavior of *the person*. The latter are those representations that play a role in explanations of the behavior of *cognitive mechanisms* of the person.³¹ Is there any reason to think that the person-level/sub-personal-level distinction maps onto the consciousness/unconsciousness distinction? Many philosophers and psychologists think so. There is widespread consensus that consciousness plays the *integration role*: it coordinates the functioning of otherwise self-contained cognitive modules, by integrating the information processed by these modules into a central, domain-general stream of information. If some representational state of an agent is not part of such a “global workspace” (to use Baars’ [1997] influential metaphor), then behavior motivated by that state does not satisfy the Control condition, because it is not controlled *by the agent*, as such, but merely by her sub-personal mechanisms. Nor would it satisfy the Identification condition: it is presumably an agent’s person-level beliefs and desires that could count as her established commitments and values. Levy (2014) argues, further, that behavior motivated by a state not part of the global workspace would *eo ipso* fail to satisfy the Reasons-Responsiveness condition as well: if a behavior is not caused by a state that is part of the global workspace, then that behavior reflects only a subset of the agent’s reasons. But for *the agent herself* to exhibit a suitable degree of reasons-responsiveness, her actions need to be responsive to the full range of her person-level mental states.³²

There is, however, room to doubt that consciousness' playing the integration-role entails that intentions governing controlled action need to be (actually or potentially) conscious. First of all, it is not clear whether consciousness *uniquely* plays the integration role. There is certainly conceptual room for a dissociation between integration and phenomenal consciousness; the two concepts are not the same. And there is some evidence of actual dissociation.³³ Second, even if consciousness plays an integration-role, it might not integrate the right kinds of items. For example, King and Carruthers (2012) contend that the contents of the global workspace include mental imagery and inner speech, but never include propositional attitudes; if this is right, then the intentions that govern controlled behavior are never conscious.

Nor is it unanimous that consciousness plays the integration-role; some theories of consciousness fail to attribute such a role to consciousness. For example, according to the "Higher-Order Thought" (or "HOT") theory of consciousness, what it is for a mental state to be conscious is for a subject to have a higher-order mental state directed at the first one. So, even if all person-level states are integrated into the global workspace, HOT-theorists deny that person-level states are therefore *conscious*. Though it is consistent with HOT theory to insist that all person-level representations are potentially conscious (that is, to insist that a subject can have a higher-order thought about all of her person-level representations), the view itself doesn't motivate this requirement. Thus, for a HOT theorist, nothing rules out the possibility that controlled actions can be governed by intentions that the agent is not capable of consciously entertaining.³⁴

There remains the final item on our list of proposed cognitive roles for consciousness, namely, that the cognitive role of consciousness is "a grab-gab." The possibility exists that one or more of the individual functions in the grab-bag is necessarily involved in the formation and/or execution of intentions. For example, David Hodgson (2012) argues that mental states he calls "gestalts" – namely, the binding together of many features in an experiential whole – can only be entertained consciously, and play a crucial role in action-planning.³⁵ Or again, Roy Baumeister (2010) presents evidence that we can only understand certain types of sequentially-structured thoughts consciously, e.g., those that depict logical and causal relations. Paradigms of such sequentially-structured thoughts are *narratives*, and it is only via the mental rehearsal of narratives that we freely choose between alternative action-plans.

There are plenty of other specific functions that consciousness is claimed to uniquely perform: constructing a self-concept; learning new, complex behaviors; building associations between events occurring at a temporal gap from one another; semantically processing sentences; planning sequences of actions, as a means to an end; thinking through a multi-step problem; and so on.³⁶ If any of these functions makes a crucial difference to the formation or execution of the intentions that govern controlled action, then consciousness will turn out to be integral to free action. But it depends on what it is for consciousness to "uniquely" perform these functions – namely, whether consciousness *necessarily* performs them, or whether it is a fluke of our evolutionary history that consciousness always or typically performs them. That human evolution delegated certain functions to consciousness doesn't mean that it had to.³⁷ Granting that humans actually rely on consciousness when they act freely, it does not immediately follow that free action and consciousness are related in any modally strong way.

3 Empirical Considerations

In recent decades, an enormous empirical literature has been growing regarding the relationship between consciousness and action in general, and consciousness and free action in particular. In much of this literature, the empirical findings are understood as threat-posing: they

challenge cherished assumptions about the relationship between our conscious states and our actions. Such challenges can be taken as having either of two implications for a connection between consciousness and free action: (a) a conceptually or metaphysically necessary connection thought to hold between consciousness and free action turns out not to hold after all; or (b) such a connection *does* hold, but fewer of our actions are actually suitably connected to our conscious states than pre-theoretic intuition would suggest (and thus we are less free than we think we are). I will discuss empirical findings from four contexts: pathologies of action; vision-for-action processing; laboratory dissociations; and situational studies.

1. *Pathologies of action.* There are a number of documented pathologies in which a subject's ordinary experience of agency is dramatically inhibited or altered.

- (1) **Anarchic Hand syndrome:** one of an agent's hands engages in apparently goal-directed behavior, sometimes at odds with what she is consciously trying to do, yet the agent is unable to stop or redirect the behavior.³⁸
- (2) **Utilization behavior:** an agent cannot prevent herself from responding habitually to an environmental stimulus.³⁹
- (3) **Delusions of alien control:** an agent experiences that and/or avows that her behaviors are under the control of an external will.⁴⁰
- (4) **Global automatism:** an agent engages in complex behaviors while unconscious (e.g., during sleep, seizure, or coma).⁴¹

In what sense do these cases challenge the idea that free will and consciousness are connected (in concept or in fact)? A natural, first reaction is that they pose no such challenge: if they are cases of action at all, they are cases of frustratingly *unfree* action, precisely because conscious oversight is diminished or absent. But, at the same time, they hint at a challenge regarding the actual extent of the connection. If in pathological cases an agent's hand can engage in complex, goal-directed behavior without or even contrary to the agent's conscious will, then perhaps one's conscious will is causally irrelevant to the behavior of one's hand even in normal cases. Behaviors as simple as picking up a cup and as complex as driving⁴² or composing and sending emails⁴³ can occur in the absence of conscious awareness, driven by non-conscious "action scripts."⁴⁴ Given that such behaviors can occur in the absence of a conscious cause, it is plausible that they occur without being consciously caused even in cases where subjects are consciously aware of performing them.

Now, there is a general consensus that pathological actions are *routinized*, in the sense that they do not express genuine novelty, creativity, or problem-solving on the part of the agent.⁴⁵ They are thus of a piece with habitual actions, as discussed above (such as squeezing the steering wheel or signaling a turn). So, the possibility of pathological action puts pressure on the constitution view and the strong causal-dependence view only to the extent that habitual action does.

2. *Functional dissociation between perceptual and motor processing.* The human visual system serves two primary functions: *acquiring information* about the local environment and *facilitating action* vis-à-vis the local environment. In the last twenty years, neuroscientists have come to believe that the processes that serve these two functions are anatomically and causally isolated from one another to a certain extent.⁴⁶ Evidence has come in the form of pathological dissociations (patients who cannot recognize objects but can act on them normally, and vice versa) as well as dissociations in healthy subjects. For example, Haffenden and Goodale (1998) used plastic discs to create a version of the Ebbinghaus Illusion, where two identical circles appear to differ in size, because one is surrounded by a ring of larger circles and the

other is surrounded by a ring of smaller circles. They found that when subjects use their thumb and index finger to estimate the size of each disk, grip aperture reflects illusory size, whereas when subjects use their thumb and index finger to *reach for* each disk, grip aperture reflects actual size.

Some theorists have interpreted these findings as indicating that a lot of motor action is “zombie action,” i.e. not consciously controlled.⁴⁷ Presumably, when subjects in the Haffenden and Goodale study use their thumb and forefinger to estimate disc size, their estimates correspond to, and are controlled by, their perceptual experience of the discs. But their actual movement vis-à-vis the disks do not correspond to, and so are not controlled by, their perceptual experience of the discs. And there is no reason to think that the illusory nature of perceptual experience in the case in question *creates* the dissociation; rather it *reveals* a dissociation that’s always at play.

But it is at best misleading to characterize subjects’ grip aperture as “zombie action.” While it’s true that *something* is unconsciously controlled when a normal subject picks up an object, that something is not an intentional action but rather a component, aspect, or specification of an intentional action, namely, the precise grip aperture used (and countless other motor details, presumably). But there is a big difference between zombie action in *that* sense and zombie action in the sense of anarchic hand movement. In the former but not the latter case, it makes sense to say that the subject is doing what she consciously intends to do.

What the Haffenden and Goodale study and others like it teach us is that much of the implementation-details of our actions are unconsciously controlled. This point is also familiar from studies of *expert* behavior, which have shown that the better an agent is at some task, the fewer of the details of the means and mechanisms whereby the agent completes the task are consciously experienced.⁴⁸ For example, when a novice pianist plays “Für Elise,” he’ll need to consciously intend to curl his fingers thus and so, to read one measure ahead, etc. An expert pianist may need to consciously intend nothing more than *to play “Für Elise.”* Reaching-behavior is, after all, just another type of (ubiquitous) expert behavior. We expert reachers need consciously intend nothing more than to *pick up the disc*; all of implementation-details are taken care of unconsciously.⁴⁹

Expert action is much like habitual action in that both are typically implemented unconsciously. But whereas habitual action is typically *triggered* unconsciously, expert action is more likely to be triggered consciously – at least under some description (*play the piece, score a point, take a drink*, etc.). That we can intend to perform actions without being consciously aware of the motor-details is not to say that those actions are not governed by conscious intentions, however much those intentions under-specify the motor-details.

3. *Laboratory dissociations.* A lively empirical research program has come together in the last three decades exploring the relationship between physical action and the phenomenology of willing.⁵⁰ Experiments of two sorts have been used to argue that actions are not, or are not always, dependent on agents’ conscious will in the way agents take them to be. First, the work of Benjamin Libet and colleagues suggests that actions can be caused by neural events that temporally precede the agent’s conscious experience of willing them. The experiments that suggest this conclusion involve instructing subjects to spontaneously flex their hand whenever they wish, and then to report the time at which they had the conscious “urge or intention” to do so. (A clock whose “second hand” – actually an arc of illuminated dots – completes a circuit in two and a half seconds is used to mark time.) Meanwhile, measurements of electromagnetic activity of the brain are taken. Libet (1985) found that a particular spike in electrical activity was consistently recorded about 350 microseconds prior to any conscious experience of willing to flex. This pre-motor spike in the electrical output of the

Supplementary Motor Area had been previously discovered and dubbed the “readiness potential.” Libet was able to show that the readiness potential appears not only before motor actions but before any experience of will associated with those actions.⁵¹

What is the significance of this result for the relationship between consciousness and freedom? Many scientists, and a few philosophers, take Libet’s experiments to have shown that we have much less free will than we thought we had. For example, Gregg Caruso has concluded, “Our conscious selves are not in the driver’s seat as we typically believe. This is a devastating blow to our everyday understanding of free will.”⁵² Those like Caruso who draw this skeptical lesson from Libet’s experiments are committed to three claims: first, that the metaphysical connection between consciousness and free action is at least as intimate as that specified by the strong causal-dependence view; second, that Libet’s experiments show that in the cases in question, consciousness plays no causal role in producing action; third, that the cases in question involve the sorts of actions we would pre-theoretically take to be paradigms of free action. But all three of these claims have been questioned. Max Velmans (1991) rejects the first, maintaining instead that from Libet’s experiments we learn that free actions can be unconsciously caused. (Libet has thus provided an empirical corrective to a conceptual error, according to Velmans.) Alfred Mele (2009) rejects the second. All that Libet’s experiments show is that something (the readiness potential) in fact occurs in the brain prior to subjects’ flexing. But this does not by itself establish that readiness potentials are sufficient causes of flexings: (a) consciousness could be contributing an intervening cause between readiness potential and flexing; (b) readiness potentials sometimes occur in the absence of any flexing, as one of Libet’s experiments demonstrates;⁵³ and (c) a conscious, freely formed, *standing* intention to flex-when-one-feels-like-it might be causally necessary.⁵⁴ And the third assumption is rejected by those who note the peculiar *pointlessness* of wrist-flexing: the subject has no reason to flex rather than not to flex at any particular moment.⁵⁵ Such “Buridan’s ass” cases – cases of choosing arbitrarily between two inconsequential options – are not very much like cases for which we take free will to be important: cases of bringing one’s cherished values to bear in deliberating about consequential options, and then directing one’s behavior accordingly. Those who grant the first and second assumptions can still insist that, if Libet has shown that we are less free than we thought we were, he has not shown that we are *very much* less free than we thought we were, or that we are not free when we *care* to be free.

There is a second type of experiment that has been used to argue for a dissociation between consciousness and free action. Daniel Wegner and colleagues have tried to show that subjects can have the experience as of willing an action without performing or even trying to perform that action – that there can be “false positives” among conscious volitional episodes. Two experiments of Wegner’s suggest this possibility. (1) The “I-Spy” experiment:⁵⁶ an experimental subject sits adjacent to a confederate of the experimenter. Both place their hands on a computer mouse that controls a cursor on a visible computer screen. The screen depicts an assortment of objects. The subject is instructed to move the cursor, or to allow the cursor to move, among the depicted objects, finally resting on any one of them after a few seconds. Meanwhile the subject hears music combined with a stream of apparently random words, which in some trials includes the name of an object depicted on the screen, and in some trials does not. For some trials, the confederate is instructed to guide the cursor to the image of the object named in the word-stream. After each trial, the subject is asked to rate the degree to which she or he had intended (vs. merely *allowed*) the cursor to land where it landed. On trials when the confederate guides the cursor to an object named in the word-stream, subjects report a higher degree of intendedness than on similar trials in which the object is not named. (Notably, the naming of a depicted object in the word-stream did not raise the probability that the subject would guide the

cursor to part of the screen where that object is depicted.) (2) The “Helping Hands” experiment:⁵⁷ an experimental subject faces a mirror, while wearing a smock that allows the arms of a confederate (who is standing directly behind the subject) to appear where his or her arms would ordinarily be. The confederate proceeds to perform a series of familiar tasks, such as clapping and making a fist. During some trials, subjects hear instructions that precede and correspond to the actions performed. After these trials, subjects reported a higher degree of intendedness than after those trials during which no instructions were audible.

Wegner (2002) concludes from these experiments, in tandem with others that purport to show the possibility of conscious-volitional “false negatives,”⁵⁸ that the conscious experience of will is always, or usually, an epiphenomenon. This inference is dubious, for a couple of reasons. First, it isn’t clear that in the “false positives” experiments, subjects have had an experience of willing at all. Rather, it seems just as likely that subjects have had no such experience, but are nevertheless inclined to attribute an action to themselves, as we often do in the absence of volitional experience (e.g., a glass is knocked off of a table; before there is any time for me to consciously will to act, I instinctively reach out to catch it – an action I happily attribute to myself), and that such attributions are fallible. Thus, Wegner has not demonstrated that conscious willing is *ever* epiphenomenal. But even if he is right that experimental subjects are having epiphenomenal conscious-volitional episodes in the I-Spy and Helping Hands experiments, his generalization to all or most conscious-volitional episodes is probably not warranted. Wegner seems to be reasoning as follows: in some cases, the experience of conscious willing is non-veridical; we ought to explain action-production in cases of non-veridical conscious willing in the same way as in cases of veridical conscious willing; so conscious willing plays no causal role in either case. But as multiple commentators have pointed out,⁵⁹ just as perceptual hallucinations do not show us that normal perception is illusory, so volitional “hallucinations” do not show us that, in normal cases, our experience of ourselves as conscious agents is illusory.

How do Wegner’s findings bear on the connection between consciousness and freedom? First, if Wegner is right that conscious, volitional episodes of a certain type – specifically, experiences of choosing to execute a bodily movement – *never* generate (let alone constitute) action, then either (a) the constitution view and the strong causal-dependence view are false or (b) one or the other is true and our actions are unfree. But the *weak* causal-dependence view is consistent with epiphenomenalism regarding consciously willing to move one’s body, so long as conscious episodes of a different sort – the conscious cultivation of a habit, the conscious experience of willing an action sometime in the future, etc. – are causally relevant to action-production. Second, if Wegner is right that the experience of willing *sometimes* fails to initiate action in any sense, then this would put pressure on the constitution view. (While advocates of the constitution view do allow that executive consciousness can fail to issue in any bodily movement, Wegner’s claim is stronger: agents can experience themselves as initiating action when they are not even *trying* to initiate action.)

4. *Situational factors.* A challenge of a different kind comes for the voluminous “situationist” literature in psychology. The studies that comprise this literature suggest that human behavior is influenced, in profound, predictable, and irrational ways, by the situations we find ourselves in. Behavior-manipulating factors described in this literature include the following:

- (1) **Bystanders:** The likelihood that a subject offers help to someone in need of medical attention is inversely proportional to the number of other people in the area, and proportional to their level of concern for the emergency.⁶⁰

- (2) **Haste:** The likelihood that a subject offers help to someone in need of medical attention is proportional to how big of a rush he or she is in.⁶¹
- (3) **Luck:** The likelihood that a subject offers help to someone who has dropped his papers is greatly increased if the subject has recently found a coin.⁶²
- (4) **Authority:** Despite expressing strong misgivings, subjects are willing to apply intense electrical shocks to someone if an authority figure is telling them to.⁶³
- (5) **Watching eyes:** Subjects contribute considerably more money to a cause when an image of a pair of watching eyes is nearby.⁶⁴

Despite the differences between the manner and extent of effect that these factors have on human behavior, a few philosophers have taken them to jointly pose a threat to free will, or at least to constrain the extent of our free will. But exactly what this threat comes to is a matter of ongoing debate.⁶⁵ In fact, different situational factors are problematic in different ways, depending on which of the four conditions on freedom is being discussed.⁶⁶

Whether or not the studies mentioned above give us reason to doubt the *extent* of our freedom, do they tell us anything about the relationship between consciousness and freedom? That is, do they demonstrate that consciousness and action are separable in ways that are at odds with the constitution view, the causal-dependence view, or the counterfactual-dependence view? They do not seem to. Consider the constitution and causal-dependence views. The influence of situational factors on human action is consistent with conscious intentions' directly contributing to action. Plausibly, when a subject puts more money in an "honesty box"⁶⁷ than she would have had an image of watching eyes not been posted nearby, or when she more readily helps a stranger than she would have had she not just discovered a dime in a phone booth, these factors influence her precisely *via* consciousness: they influence the degree of her *felt inclination* toward performing some action. So long as it is still in some sense "up to her" whether to follow that inclination, she is still exerting conscious control over her actions.

The counterfactual-dependence view seems to receive some confirmation from the situationist literature. Consider a version of the counterfactual-dependence view according to which an action is free only if (a) the agent can consciously entertain the reasons for which she acts, and (b) can consciously entertain reasons for acting otherwise. When subjects are under the influence of the bystander effect, it's plausible that they act out of a desire to "fit in" with the social group. Yet these same subjects explicitly deny being influenced by the presence and behavior of the other people in the study.⁶⁸ In these circumstances subjects *cannot* consciously entertain (some of) the reasons for which they act, contra clause (a). When subjects are under the influence of the haste effect, it's plausible that they are acting without properly considering the needs of the "medical patient" whose path they cross. In these circumstances subjects cannot consciously entertain (some) reasons for acting otherwise than they do, contra clause (b). The counterfactual-dependence view is thus able to explain in what sense subjects under the influence of bystander and luck effects are less than fully free.

Whether an agent *has the capacity* in a situation to consciously entertain certain reasons is not straightforward; it depends on which aspects of the situation are held fixed. For example, I may be capable of consciously entertaining a reason for acting *if only I would pause for a moment to reflect*, or *if only I were informed about the situationist literature in social psychology*,⁶⁹ or *if only I had received a more rigorous moral education*. Thus the question "Was action A done freely"? will be answerable in many cases only relative to a set of possibilities, and in terms of a notion of freedom that can come in degrees.⁷⁰

I have reviewed four of the most frequently discussed empirical findings related to the connection between consciousness and action, but there are others. Some psychologists have

taken these findings to present a radically revisionary conception of human behavior, one that demotes conscious processing to a mere spectatorial role, and that gives pride of place to unconscious processes.⁷¹ But the growing consensus is that these revisionary conclusions are not warranted by the experimental data.⁷² Moreover, empirical evidence is accumulating that supports the opposite conclusion. Studies on the causal efficacy of consciousness have revealed that consciousness regularly drives action in precisely the ways we intuitively expect, such as via deliberation and planning (Baumeister, Masicampo, and Vohs 2011), and even in moment-to-moment action control, when rival motivations and action-plans present themselves (e.g., when the impulse to drop a hot plate conflicts with the goal of carrying it to the dinner table) (Morsella 2005). Far from undercutting a conception of ourselves as in conscious control of our actions, empirical science seems to be confirming it.

4 Conclusion

Though very few of the considerations we have explored decisively favor one view over its rivals, nevertheless the following two lessons can reasonably be drawn from the foregoing.

First, the possibility of free habitual and expert actions gives us good reason to think that the constitution and strong causal-dependence views are too strong. When philosophers and scientists have taken the empirical discoveries of Libet and Wegner as a threat to free will, they have implicitly assumed that all free actions must be constituted or directly caused by consciousness. Since that assumption is too strong, we can rule out the more revisionary of glosses on Libet and Wegner's work. But this doesn't mean that the constitution and causal-dependence views are without merit. They rest on the intuitive idea that the *direct exercise of control* is a conscious matter. The weak causal dependence view can accommodate this intuition. For example, it could be held that every free action is either a constituted by consciousness (e.g., by being identical to a conscious trying) or is the causal descendent of an action constituted by consciousness. This is a plausible way of understanding what the Control condition requires.

Second, if the Control condition requires that free action is causally dependent on consciousness, the Reasons-Responsiveness condition plausibly requires that free action is counterfactually dependent on consciousness. One strong reason to think so is that it explains how the situationist literature in cognitive psychology exposes limitations to our freedom. Even if our actions trace their causal ancestry to the conscious exercise of agentic control, if we are incapable – at least holding fixed certain aspects of the context – of consciously entertaining the reasons for which we act, then we do not act freely, in the fullest sense. Free will is probably conceptually connected to consciousness in more than one way – causally, in terms of control, and counterfactually, in terms of reasons-responsiveness.⁷³

Notes

- 1 This is a classic, if currently highly contested, view of the epistemic role of consciousness. It goes back at least to Descartes and the British empiricists. See Gertler (2012) for discussion of some recent re-treads.
- 2 John Searle (1991) is responsible for putting this view on the table.
- 3 Bernard Baars' (1997) "Global Workspace Theory" of consciousness is the classic version of this view, but it has been developed in many ways.
- 4 See e.g. Marcel (1988), Hurley (1997), Baumeister, Masicampo and Vohs (2011), and Dehaene (2014).

- 5 Reasons-Responsiveness is sometimes treated as an account of control, and sometimes as a separate condition.
- 6 The debate between compatibilists and incompatibilists turns on whether the Open-Future condition is included in our concept of freedom. I will mention the Open-Future condition only very briefly in what follows; controversies surrounding the other three conditions matter more to whether and how consciousness and freedom are connected. Thus the question of how consciousness and freedom are related is largely independent of the debate between compatibilists and incompatibilists.
- 7 The assertion that phenomenal-consciousness and access-consciousness are close-enough-to-coextensional for present purposes calls for a more thorough defense than I can give here. As a start: most consciousness-researchers seem to assume that, under most circumstances, mental items are phenomenally conscious if and only if they are available for verbal report. There are two reasons to think so. First, at least until quite recently, researchers in philosophy and psychology overwhelmingly claimed phenomenal consciousness as the target of their inquiry, even as Block (1995) told them they were conflating two notions. Second, excepting the study of unusual psychological effects (blindsight, subthreshold priming, etc.), phenomenal- and access-consciousness are experimentally operationalized in exactly the same way, namely, via verbal report. When a subject is unable to report being aware of some stimulus (or explicitly reports being unaware), this is taken as an indicator that she is *unconscious* of that stimulus, not that it is conscious in some non-phenomenal way.
 My own view is that access-consciousness, as Block originally delineated the category, shouldn't be treated as a species of consciousness. Nevertheless, there is a related category of non-phenomenal, occurrent mental states that should be studied alongside phenomenal consciousness. This is the category of the states that are phenomenally conscious or waiting-in-the-wings to become so (e.g., representations of items on a grocery list that one has committed to working memory). I suspect that consciousness-researchers who say they are studying access-consciousness are often studying this phenomenology-*plus* type of consciousness.
- 8 Horgan, Tienson and Graham (2003), Bayne (2008), and Nida-Rümelin (2017).
- 9 That there is irreducibly executive phenomenology is an outlier view in contemporary analytic philosophy and cognitive science, but seems to have been the standard view in nineteenth- and twentieth-century phenomenology. See Kriegel (2015), p. 83. In the eighteenth century Thomas Reid wrote that "We are conscious of making an exertion, sometimes with difficulty, in order to produce certain effects" (1788/1983, p. 332).
- 10 That is, there may be many things that an agent is trying to do by a single action. But only one of these things she is trying to do is identical to her *basic* action, i.e., the action the agent performs without performing any other action as a means to performing it. Danto (1963) first drew attention to the category of basic action.
- 11 For a recent discussion of this "argument from total failure" see Grunbaum (2008). It's noteworthy that David Hume discusses a total-failure case, but reaches a negative conclusion about executive phenomenology: "A man suddenly struck with a palsy in the leg or arm, or who had newly lost these members, frequently endeavours, at first, to move them, and employ them in their usual offices. Here he is as much conscious of power to command such limbs, as a man in perfect health is conscious of power to actuate any member which remains in its natural state and condition. But consciousness never deceives. Consequently, neither in the one case nor in the other, are we ever conscious of any power" (1777/1975, p. 66).
- 12 O'Shaughnessy (1973), p. 370. Hossack (2003) talks of "volitions" rather than "tryings," but seems to have the same phenomenon in mind.
- 13 Compare Schroeter (2004): "I ... suggest that the agent is the conscious self" (p. 642); a conscious self "has the power to send go-signals to initiate the execution of the actions which have been selected" (245). (Note, though, that Schroeter says he is talking about the *access-conscious* self. And whether he endorses the constitution view or merely the causal-dependence view is not clear.)
- 14 O'Shaughnessy (2003), O'Brien (2003), and Hossack (2003) all appeal to versions of the Self-Awareness Argument.

- 15 O'Connor (1995).
- 16 Chalmers and McQueen (Forthcoming) have also developed a wave-collapse theory of mental causation, though not explicitly as an account of libertarian free will.
- 17 Arthur Danto (1985, p. 541) makes much the same point in the context of *projects* rather than *habits*: "A slow-motion film of Matisse shows the artist making countless decisions with his fingers that at normal speed looks like a single confident chalk stroke defining the edge of a leaf. He may or may not have been conscious of each decision ... Consciousness is evolution's gift to us for rather special deliberative employment having to do, as responsibility and free will have to do, with *courses* of action – with *projects*."
- 18 This example comes from Mele (2009).
- 19 From their study of folk conceptions of free action, Stillman, Baumeister, and Mele (2011) conclude that freedom is generally associated with conscious deliberation about, and successful pursuit of, long-term goals. Thus the Indirect Control condition probably better reflects the folk notion of free action than does the Direct Control condition. In keeping with this theme, Nahmias (2002, p. 537) writes, "We all aim to be able, in many types of situations, to act without the feeling of will, but in accord with the plans we have consciously formed. And we feel free and responsible for the actions that accord with these plans, regardless of whether we consciously will each of them."
- 20 Fischer and Ravizza (1998) label (roughly) this type of counterfactualist reasons-responsiveness "guidance control."
- 21 See Searle (1991) p. 56.
- 22 See McGinn (1988), Searle (1991), and Kriegel (2003).
- 23 This problem, sometimes labelled "the disjunction problem," is carefully developed in Fodor (1990). It has close affinities with the "gavagai" problem discussed at length by Quine in his (1960). Fodor thinks the problem can be solved; others are not so optimistic, such as Searle (1991), Adams and Aizawa (1994), and Horgan and Graham (2012).
- 24 Searle (1991), Ludwig (1996) and Strawson (2008) all give versions of this argument.
- 25 As it happens, King and Carruthers (2012) argue that no propositional attitudes are ever conscious, and thus that there probably is no such thing as a "true self" in the sense presupposed by the Identification condition.
- 26 Relatedly, Neil Levy (2014) argues that there is a necessary connection between access-consciousness and moral responsibility: in order to be morally responsible for their actions, agents must be conscious of facts that explain the valence of [their actions'] moral significance" (2014, p. 37). For those who think (as I do, but Levy does not) that an access-conscious state is "conscious" in any meaningful sense only if it is a potentially phenomenally conscious state, Levy's view counts as a version of the counterfactual-dependence view.
- 27 See, e.g. Ruiz-Mirazo and Moreno (2004) and Jones (2015).
- 28 See Jeannerod (1997).
- 29 Schroeter (2004) and Shepherd (2015) both stress this role for consciousness.
- 30 Of course, we hesitate to categorize somnambulant behaviors as free actions. But if the comparator model is right, this can't be because the somnambulist fails to satisfy the Control-condition.
- 31 See Drayson (2014) for a helpful history of the personal/subpersonal distinction in cognitive science.
- 32 It's not clear to me whether Levy is thus committed to the causal-dependence view or the counterfactual-dependence view. That is, it is not clear whether he thinks that behavior that satisfies the Reasons-Responsiveness condition has to issue from *actually* or merely *potentially* globally-available mental states.
- 33 Morsella (2005). Levy (2014) reviews and argues against a number of recent, putative cases of dissociation.
- 34 David Rosenthal, the original and longest-standing advocate of HOT theory, explicitly embraces the epiphenomenality of consciousness when it comes to action-explanation in his (2008). As far as I can tell he is silent on the question of whether controlled action can be guided by intentions the agent is not even capable of consciously entertaining (i.e. whether the counterfactual-dependence view is true).

- 35 Baumeister (2010) seems to concur: “Consciousness is the picture the brain constructs for itself from a (nearly) impossible number of individual pieces of sensory information” (40).
- 36 The literature is voluminous, growing, and rich with controversy. See Marcel (1988), Hurley (1997), Baumeister, Masicampo and Vohs (2011), and Dehaene (2014).
- 37 Polger (2007).
- 38 Marchetti and Della Sala (1998).
- 39 Lhermite (1983).
- 40 Spence (2002).
- 41 Cartwright (2004), Levy and Bayne (2004).
- 42 Penfield (1975).
- 43 Siddiqui, Osuna and Chokroverty (2009).
- 44 See Levy (2014), p. 74ff.
- 45 Levy (2014), Dehaene (2014).
- 46 Milner and Goodale (1995). Milner & Goodale’s original picture was probably too tidy; see Shepherd (2015).
- 47 Cf. Clark (2001), Koch and Crick (2001), Wu (2013).
- 48 Shepherd (2015).
- 49 I am granting, along with the majority of philosophers of cognitive science, that the Haffenden and Goodale study and others like it reveal that there are unconsciously controlled aspects of expert action. See Ian Phillips (2018) for an important dissenting opinion. He argues that the relevant experiments do not warrant even this modest conclusion.
- 50 The essays in Pockett, Banks, and Gallagher (2006) amount to opinionated summaries of much of this research.
- 51 Many Libet-style experiments have been performed in the past three decades, some apparently supporting Libet’s original conclusions, and some calling those conclusions into serious question. Haggard and Eimer (1999) claim to have shown that the experience of willing is regularly preceded not by the readiness potential but by a different event, the “lateralized readiness potential.” Schlegel et al. (2013) partly replicate these findings, but do not think they license Libet’s conclusions: “Neither the RP nor LRP appears to be a correlate of consciousness or conscious volition. The RP may instead be a correlate of general anticipation or preparation or merely of ongoing activity that is neither anticipator nor preparatory” (p. 334) Schurger, Sitt and Dehaene (2012) likewise suggest that the RP be treated as so much neural noise.
- Soon et al.’s (2008, 2013) experiments include a choice between two rival actions – pushing buttons with one or the other index finger (2008); adding vs. subtracting numbers (2013). They find that the readiness potential is not the earliest predictor of which choice subjects will make; as early as 7 seconds prior to motor action (2008) and 4 seconds prior to mental action (2013), neural activity in frontopolar cortex and parietal cortex occurs that can be used to predict which action the subject will perform – but only with about 60% accuracy.
- A number of commentators have from the very beginning raised questions about the design of Libet’s experiments (see peer commentary included in Libet [1985]), but for present purposes I am granting that Libet’s experimental paradigm is basically sound. For much more on Libet-style experiments and their interpretation, see Sinnott-Armstrong and Nadel (2010).
- 52 Cf. Caruso (2012), p. 194. By “understanding of free will” Caruso appears to mean our *beliefs about the extent* of free will, rather than our conception of it.
- 53 In one of the experiments, subjects were instructed to (a) *intend* to flex, but (b) consciously *veto* this intention. Libet reports that “A large RP [readiness potential] preceded the veto, signifying that the subject was indeed *preparing* to act, even though the action was aborted by the subject” (Libet 1999, p. 52). Thus Libet himself rejects an inference from his findings to epiphenomenalism. Now, Mele’s gloss is that Libet’s instructions (a) and (b) cannot be followed simultaneously, and thus that Libet interprets the results incorrectly. Either way, Libet’s data include cases of readiness potentials that are not followed by flexings.

- 54 Eiler and Roessler (2003) make much the same point: “the causally relevant contribution of the subjects’ conscious intentions lies in their prior decision to move their hand “sometime,” delegating the task of deciding exactly when to move to some subpersonal mechanism...Under these circumstances, delegating control to subpersonal mechanisms may be precisely a way of choosing when to act” (42).
- 55 See Breitmeyer (1985), p. 539; Mele (2009), p. 85.
- 56 Wegner and Wheatley (1999).
- 57 Wegner describes the experiment in his (2002), pp. 80–81.
- 58 I.e., cases in which an agent acts but has no conscious experience of doing so. Wegner calls these cases “automatisms.” They include “table-turning,” a practice popular in the nineteenth century in which a group of people lay hands on a table-top which proceeds to rotate, and this motion is attributed to occult forces; and “facilitated communication,” in which a helper “assists” a disabled person to type meaningful prose.
- 59 E.g. Nahmias (2002), Bayne (2006), and Shepherd (2015).
- 60 Latane and Darley (1968).
- 61 Darley and Batson (1973).
- 62 Isen and Levin (1972). This study remains influential despite concerns about replicability. Cf. Blevens and Murphy (1974) and Weyent and Clark (1976).
- 63 Milgram (1974).
- 64 Bateson, Nettle and Roberts (2006).
- 65 See Nelkin (2005), Vargas (2013), and Herdova (2016).
- 66 Nelkin (2005) and Vargas (2013) make a strong case that the central threat is to Reasons-Responsiveness. For example, it is plausible that haste prevents subjects from responding to a need precisely because haste prevents them from being able to engage in the sort of deliberation required (a) to recognize the need as a reason for acting and/or (b) to act for the sake of that reason once recognized. But not all of the factors are troubling for this reason. Luck – e.g., finding a dime in a phone booth – seems actually to boost subjects’ capacity to recognize a need as a reason to act. I think the best way to characterize the perceived threat to freedom from situational luck is that it falsifies the Control-condition: the subject’s decision to help appears to be dictated by her elevated mood, and so no more up to her than other blessings of an elevated mood (increased focus, optimism, etc.). The influence of authority in Stanley Milgram’s experiments is probably best understood as a falsification of the Identification-condition, since subjects appear to act contrary to their best judgment. None of the factors can be seen as a falsification of the Open Future condition, since in the relevant experiments, subjects’ behavior fails to be fully predictable given situational factors.
- 67 That is, the designated cash receptacle for purchased made “on the honor system,” i.e. in the absence of a cashier.
- 68 “Despite the obvious and powerful inhibiting effect of other bystanders, subjects almost invariably claimed that they had paid little or no attention to the reactions of the other people in the room.” (Latané and Darley 1968, p. 220).
- 69 Mele and Shepherd (2013) argue that becoming familiar with situationist literature is a way of increasing one’s degree of free agency.
- 70 Vargas (2013) argues at length for this idea.
- 71 See e.g. Bargh and Ferguson (2000) and Dijksterhuis, Chartrand and Aarts (2007).
- 72 Newell and Shanks (2014), Baumeister, Masicampo and Vohs (2011).
- 73 Thanks to Neil Levy, Sandra Visser, Tim O’Connor, and Lisa Woodward for helpful comments on an earlier draft of this chapter.

Bibliography

- Adams, F. and Aizawa, K. (1994). Fodorian semantics. In: *Mental Representation* (ed. S. Stich and T. Warfield). 223–242, Cambridge, MA: Blackwell.
- Baars, B. (1997). *In the Theater of the Consciousness: The Workspace of the Mind*. Oxford: Oxford University Press.

- Bargh, J. and Ferguson, M. (2000). Beyond behaviorism: on the automaticity of higher mental processes. *Psychological Bulletin* 6: 925–945.
- Bateson, M., Nettle, D., and Roberts, G. (2006). Cues of being watched enhance cooperation in a real-world setting. *Biology Letters* 2: 412–414.
- Baumeister, R. (2010). Understanding free will and consciousness on the basis of current research findings in psychology. In: *Free Will and Consciousness: How Might They Work?* (ed. R. Baumeister, A. Mele and K. Vohs), 24–42. Oxford: Oxford University Press.
- Baumeister, R.F., Masicampo, E.J., and Vohs, K.D. (2011). Do conscious thoughts cause behavior? *Annual Review of Psychology* 62: 331–361.
- Bayne, T. (2006). Phenomenology and the feeling of doing: Wegner on the conscious will. *Does consciousness cause behavior*, 1, 169–186.
- Bayne, T. (2008). The phenomenology of agency. *Philosophy Compass* 3: 182–202.
- Blevens, G. and Murphy, T. (1974). Feeling good and helping: further phonebooth findings. *Psychological Reports* 34: 384–388.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18: 227–287.
- Breitmeyer, B. (1985). Problems with the psychophysics of intention. *Behavioral and Brain Sciences* 8 (4): 539–540.
- Cartwright, R. (2004). Sleepwalking violence: a sleep disorder, a legal dilemma, and a psychological challenge. *American Journal of Psychiatry*, 161(7), 1149–1158.
- Caruso, G. (2012). *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lanham, MD: Lexington Books.
- Chalmers, D. and McQueen, K. (Forthcoming). Consciousness and the collapse of the wave function. In: *Consciousness and Quantum Mechanics* (ed. S. Gao). Oxford: Oxford University Press.
- Clark, A. (2001). Visual experience and motor action: are the bonds too tight? *Philosophical Review* 110 (4): 495–519.
- Danto, A. (1963). What we can do. *Journal of Philosophy* 60 (15): 435–445.
- Danto, A. (1985). Consciousness and motor control. *Behavioral and Brain Sciences*, 8(4), 540–541.
- Darley, J. and Batson, C. (1973). From Jerusalem to Jericho: a study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology* 27 (1): 100–108.
- Dehaene, S. (2014). *Consciousness and the Brain*. New York: Penguin Books.
- Dijksterhuis, A., Chartrand, T.L., and Aarts, H. (2007). Effects of priming and perception on social behavior and goal pursuit. In: *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes* (ed. J.A. Bargh), 51–132. Psychology Press.
- Drayson, Z. (2014). The personal/subpersonal distinction. *Philosophy Compass* 9 (5): 338–346.
- Eiler, N. and Roessler, J. (2003). Agency and self-awareness: mechanisms and epistemology. In: *Agency and Self-Awareness: Issues in Philosophy and Psychology* (ed. N. Eiler and J. Roessler), 1–47. Oxford: Clarendon Press.
- Fischer, J. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
- Fodor, J. (1990). *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.
- Gertler, B. (2012). Renewed acquaintance. In: *Introspection and Consciousness* (ed. D. Smithies and D. Stoljar), 89–123. New York: Oxford University Press.
- Ginet, C. (1990). *On Action*. New York: Cambridge University Press.
- Grunbaum, T. (2008). Trying and the arguments from total failure. *Philosophia* 36: 67–86.
- Haffenden, A. and Goodale, M. (1998). The effect of pictorial illusion on prehension and perception. *Journal of Cognitive Neuroscience* 10: 122–136.
- Haggard, P. and Eimer, M. (1999). On the relation between brain potential and the awareness of voluntary movements. *Experimental Brain Research* 126: 128–133.
- Halvorson, H. (2011). *The Measure of All Things: Quantum Mechanics and the Soul*. In: *The Soul Hypothesis: Investigations into the Existence of the Soul* (ed. M.C. Baker and S. Goetz), 138–167. Continuum Press.
- Herdova, M. (2016). What you don't know can hurt you: situationism, conscious awareness, and control. *Journal of Cognition and Neuroethics* 4: 45–71.

- Hodgson, D. (2012). *Rationality + Consciousness = Free Will*. New York: Oxford University Press.
- Horgan, T. and Graham, G. (2012). Phenomenal intentionality and content determinacy. In: *Prospects for Meaning* (ed. R. Schantz), 321–344. Germany: De Gruyter.
- Horgan, T., Tienson, J., and Graham, G. (2003). The phenomenology of first-person agency. In: *Physicalism and Mental Causation* (ed. S. Walter and H.-D. Heckmann), 323–341. Imprint Academic.
- Hossack, K. (2003). Consciousness in act and action. *Phenomenology and the Cognitive Sciences* 2: 187–203.
- Hume, D. (1777/1975). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3e (ed. L.A. Selby-Bigge). Oxford: Oxford University Press.
- Hurley, S.L. (1997). Nonconceptual self-consciousness and agency: perspective and access. *Communication and Cognition* 30: 207–248.
- Isen, A. and Levin, P. (1972). Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology*, 21(3), 384.
- Jeannerod, M. (1997). *The Cognitive Neuroscience of Action*. Oxford: Blackwell.
- Jones, D. (2015). Agency through autonomy: self-producing systems and the prospect of bio-compatibilism. *Journal of Cognition and Neuroethics* 3 (1): 217–228.
- King, M. and Carruthers, P. (2012). Moral responsibility and consciousness. *Journal of Moral Philosophy* 9: 200–228.
- Koch, C. and Crick, F. (2001). On the Zombie within. *Nature* 411 (6840): 893–893.
- Kriegel, U. (2003). Is intentionality dependent upon consciousness? *Philosophical Studies* 116 (3): 271–307.
- Kriegel, U. (2015). *The Varieties of Consciousness*. New York: Oxford University Press.
- Latane, B. and Darley, J. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology* 10 (3): 215–221.
- Levy, N. (2014). *Consciousness and Moral Responsibility*. New York: Oxford University Press.
- Levy, N. and Bayne, T. (2004). Doing without deliberation: automatism, automaticity, and moral accountability. *International Review of Psychiatry* 16: 209–215.
- Lhermitte, F. (1983). ‘Utilization behavior’ and its relation to lesions of the frontal lobes. *Brain* 106 (2): 237–255.
- Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8 (4): 529–566.
- Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies* 6 (8–9): 47–57.
- Ludwig, K. (1996). Explaining why things look the way they do. In: *Perception* (ed. K. Akins), 18–60. New York: Oxford University Press.
- Marcel, A. (1988). Phenomenal experience and functionalism. In: *Consciousness in Contemporary Science* (ed. A. Marcel and E. Bisiach), 121–158. New York: Oxford University Press.
- Marchetti, C. and Della Sala, S. (1998). Disentangling the alien and anarchic hand. *Cognitive Neuropsychiatry* 3: 191–207.
- McGinn, C. (1988). Consciousness and content. *Proceedings of the British Academy* 74: 225–245.
- Mele, A. (2009). *Effective Intentions*. New York: Oxford University Press.
- Mele, A. and Shepherd, J. (2013). Situationism and agency. *Journal of Practical Ethics* 1 (1): 62–83.
- Milgram, S. (1974). *Obedience to Authority: An Experimental View*. New York: Harper & Row.
- Milner, D. and Goodale, M. (1995). *The Visual Brain in Action* Oxford: Oxford University Press.
- Morsella, E. (2005). The function of phenomenal states: supramodular interaction theory. *Psychological Review* 4: 1000–1021.
- Nahmias, E. (2002). When consciousness matters: a critical review of Daniel Wegner’s *The Illusion of Conscious Will*. *Philosophical Psychology* 15: 527–541.
- Nelkin, D.K. (2005). Freedom, responsibility and the challenge of situationism. *Midwest Studies in Philosophy* 29 (1): 181–206.
- Newell, B., and Shanks, D. (2014). Unconscious influences on decision making: A critical review. *Behavioral and brain sciences*, 37(1), 1–19.
- Nida-Rümelin, M. (2017). Freedom and the phenomenology of agency. *Erkenntnis* (83): 1–27.

- O'Brien, L. (2003). On knowing one's own actions. In: *Agency and Self-Awareness: Issues in Philosophy and Psychology* (ed. J. Roessler and N. Eilan), 358–382. New York: Clarendon Press.
- O'Connor, T. (1995). *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. New York: Oxford University Press.
- O'Shaugnessy, B. (1973). Trying (as the mental 'pineal gland'). *Journal of Philosophy* 70 (13): 365–386.
- O'Shaugnessy, B. (2003). The epistemology of physical action. In: *Agency and Self-Awareness: Issues in Philosophy and Psychology* (ed. J. Roessler and N. Eilan). New York: Clarendon Press.
- Penfield, W. (1975). *The Mystery of the Mind*. Princeton, NJ: Princeton University Press.
- Phillips, I. (2018). Unconscious perception reconsidered. *Analytic Philosophy* 59 (4): 471–514.
- Pockett, S., Banks, W.P., and Gallagher, S. (2006). *Does Consciousness Cause Behavior?* MIT.
- Polger, T. (2007). Rethinking the evolution of consciousness. In: *Companion to Consciousness* (ed. S. Schneider and M. Velmans), 77–92. Oxford: Blackwell.
- Quine, W. (1960). *Word & Object*. Cambridge, MA: MIT.
- Reid, T. (1788/1983). *Thomas Reid's Inquiries and Essays* (ed. R.E. Beanblossom and K. Lehrer). Indianapolis, IN: Hackett.
- Rosenthal, D. (2008). Consciousness and its function. *Neuropsychologia* 46: 829–840.
- Ruiz-Mirazo, K. and Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life* 10 (3): 235–259.
- Schlegel, A. et al. (2013). Barking up the wrong tree: readiness potential reflect processes independent of conscious will. *Experimental Brain Research* 229: 329–335.
- Schroeter, F. (2004). Endorsement and autonomous agency. *Philosophy and Phenomenological Research* 69: 633–659.
- Schurger, A., Sitt, J.D., and Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences* 109: E2904–2913.
- Searle, J. (1991). Consciousness, unconsciousness and intentionality. *Philosophical Issues* 1 (1): 45–66.
- Shepherd, J. (2015). Conscious control over action. *Mind and Language* 30 (3): 320–344.
- Siddiqui, F., Osuna, E., and Chokroverty, S. (2009). Writing emails as part of sleepwalking after increase in zolpidem. *Sleep Medicine* 10: 262–264.
- Sinnott-Armstrong, W. and Nadel, L. (2010). *Conscious Will and Responsibility: A Tribute to Benjamin Libet*. New York: Oxford University Press.
- Smithies, D. (2012). The mental lives of zombies. *Philosophical Perspectives* 26 (1): 343–372.
- Soon, X. et al. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience* 11 (5): 543–545.
- Soon, X. et al. (2013). Predicting free choices for abstract intentions. *Psychological and Cognitive Sciences* 110 (15): 6217–6222.
- Spence, S. (2002). Alien motor phenomena: A window on to agency. *Cognitive Neuropsychiatry* 7 (3): 211–220.
- Stapp, H. (2005). Quantum interactive dualism: an alternative to materialism. *Journal of Consciousness Studies* 12: 43–58.
- Stillman, T.F., Baumeister, R.F., and Mele, A.R. (2011). Free will in everyday life: autobiographical accounts of free and unfree actions. *Philosophical Psychology* 24: 381–394.
- Strawson, G. (2008). *Real Materialism and Other Essays*. New York: Oxford University Press.
- Vargas, M. (2013). Situationism and moral responsibility: free will in fragments. In: *Decomposing the Will* (eds. T. Vierkant, J. Kiverstein and A. Clark), 325–350. New York: Oxford University Press.
- Velmans, M. (1991). Is human information processing conscious? *Behavioral and Brain Sciences* 14 (4): 651–669.
- Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT.
- Wegner, D. and Wheatley, T. (1999). Apparent mental causation: sources of the experience of will. *American Psychologist* 54: 480–492.
- Weyant, J. and Clark, R. (1976). Dimes and helping: the other side of the coin. *Personality and Social Psychology Journal* 3: 107–110.
- Wu, W. (2013). The case for zombie agency. *Mind* 122 (485): 217–230.

Neuroscience

R.R. WALLER

The aim of this chapter is to offer a survey of the neuroscience of free will, particularly those studies of agency intimately tied to a now famous study by Benjamin Libet and colleagues. First I provide some background from action theory regarding the critical concepts of intention, decision, intentional action, and free action. I then outline Libet and company's 1983 study and its import for the development of the neuroscience of free will. Three potential threats to the existence of intentional and free action on the basis of Libet's studies are characterized. These threats are assessed via a tour of the replications and extensions of the Libet results. I then press a further objection to the broadly defined Libet paradigm, namely, that these neuroscientific studies of agency cannot generalize to conclusions about paradigmatically free decisions and actions due to the nature of actions tested in the laboratory. Using this metric of validity, I note some improvements in the Libet paradigm in the last few years. Ultimately I conclude that even given marked improvements in the kinds of action under study in neuroscience, we ought to be wary of any strong claims regarding free will on the basis of these studies.

1 Intention, Intentional Action, and Free Action

We make an extraordinary number of decisions a day, from seemingly innocuous ones like deciding to pick up that coffee cup for a sip to more important ones like my daily decision to get out of bed now to start my morning run to the occasional anxiety-inducing, life-defining ones. When I decide what to do – for instance, decide to start my run just now – I may unreflectively believe that it was up to me what I decided and, moreover, that it is up to me when and how I act – in this case when and how to move my body. Here *up to me* marks a particular sense in which I, the agent, exercise control over my decisions and actions. Philosophers often speak of these kinds of goal-directed actions as intentional actions.¹

Offering an analysis of intentional action is beyond the scope of this chapter, but for our purposes we can note that philosophers of action take intentional action to be roughly those actions characterized by an appropriate connection to our practical intentions, reasons, and/or desires and beliefs.² For instance, I might be said to run along the trail intentionally in the sense that my so running is governed by my intention to run three miles of trails this morning. That is, my starting and continuing this run is due to said intention. Here by *intention* I mean a mental state that is linked to the agent's beliefs and desires and can be characterized by an attitude of settledness toward a plan of action.³ When agents decide what to do, such as my deciding to pick up that cup of coffee, agents are forming intentions to act in certain ways.

If we believe that at least some of our decisions and actions, especially our intentional bodily movements, are up to us, we may believe that we are *freely* deciding or acting, at least at times. To decide freely or to act freely is to exercise free will, a distinctive kind of agential control. Theorists disagree about the kinds of capacities and conditions that constitute and enable agents to act (or decide) freely. One widely shared tenet, however, is that free will is the ability to act freely, and that agents who have free will have the appropriate kind of control over their decisions and actions to qualify as the type of agent deserving of moral blame (for a morally bad action) and moral praise (for a morally good action) (e.g., Watson 1996; Fischer and Ravizza 1998; Fischer and Tognazzi 2011).

Hence, when philosophers – and presumably nonphilosophers – make claims about acting intentionally and freely, they are concerned with the control actors have over their decisions and movements (and consequences of those movements). Further, it is often implicitly assumed that an agent must make some, if not all, of her decisions consciously in order to decide and act freely.⁴ This assumption underlies the focus in accounts of free will on the deliberative process leading to decision, in which agents consciously weigh reasons for performing one action versus alternatives on the basis of their background psychology.⁵

2 Libet's Seminal Study of Volition

Neuroscience of agency has made much progress in investigating the neurological underpinnings of human practical decision-making and intentional actions in the laboratory, especially endogenously generated actions. Endogenously generated actions are actions for which the cue for acting does not come from the experimenter or other environmental stimuli but rather from within the participant herself. For instance, the participant, the agent, pushes a button when or how they decide or desire to do so. Such actions in the laboratory serve as a model of self-initiated intentional actions in everyday life, such as my taking the first step of my run when I choose to do so. Early work by Kornhuber and Deeke (1965), using electroencephalography (EEG), demonstrated that when participants make simple endogenously generated movements there is a detectable slow negative buildup of brain activity in the supplementary motor area (SMA) and pre-SMA. Averaging over multiple trials (token movements) and multiple participants, this detectable increasing brain activity forms a ramp-up pattern termed the “Bereitschaftspotential” or Readiness Potential (RP) prior to movement. The RP was taken to indicate neural preparation for action.

Much of the contemporary work on neuroscience of free will can trace its lineage to a descendent paradigm of this research on the RP, the influential Libet paradigm. As we've noted, our experience as agents recommends that intentional actions are driven by intentions to act, particularly conscious intentions – those intentions we are aware of having. My taking that first step of the run is causally due, in part, to my having an intention to run just then. And the RP, plausibly, registers the body's preparation to move. Libet and his colleagues (1983), then, sought to test whether conscious intentions do, in fact, initiate or cause our intentional actions.⁶ As Libet later put the question: "When does the conscious wish or intention (to perform the act) appear?" (2011, p. 2). That is, when does the will or decision⁷ come on the scene in relation to the RP and movement? One plausible prediction, if conscious intentions to move initiate action, is that those conscious intentions occur prior to or simultaneously with neural preparation to move.

Building upon previous work, Libet, Wright, and Gleason (1982) found that the RP preceded simple endogenously generated actions, such as wrist and finger flexes. Libet and collaborators (1983) asked participants to flex their wrist or finger in each trial. In some of the trials the participants were to preplan when they would make a movement in the trial; in others the participants were instructed to flex when they felt the urge or intention to do so – i.e., spontaneously. The start of the trial was indicated by a beep. Participants watched a sped-up analogue clock on a computer screen, one that made a revolution every 2.56 seconds, while both EEG and electromyography (EMG) measurements (of movement) were recorded. After each intentional finger or wrist movement, participants reported the time of their first awareness of their intention to move using the position of the rotating dot on the clock. This reported time index is called *W*-time. Averaging over 40 trials, the average time of RP ramp up for preplanned trials was roughly a second (1025 ms) prior to recorded movement, the flexing. Averaging over 40 trials, the average time of RP ramp up for spontaneous trials was about a half second (517 ms) before recorded movement ("M"). Strikingly, in comparison, the average reported time of first awareness of intention to move, average *W*-time, occurred about 200 ms (206 ms) prior to recorded movement (*see figure 1*). Notably, Libet, with neuroscientists following suit, has emphasized the results from spontaneous trials as most reflective of the kinds of "capricious" actions and conscious decisions characteristic of intentional and free action. Granting this focus on spontaneous actions,⁸ Libet and colleagues found, then, that the RP preceded the reported time of conscious intention to move by approximately a third of a second.

Libet (Libet et al. 1983; Libet 1985) and others (Wegner 2002) infer from these results that unconscious brain processes decide (form an intention) to move prior to our awareness of our so intending; On this view, then, when we act intentionally, conscious intentions occur temporally later than the operative intention to move. This conclusion is dependent on the assumption that the RP measures neural preparation for action and that *W* is an accurate measure of first awareness of intention.⁹ Further, on the basis of his results, Libet and others conclude that conscious intentions never precede and so, presumably, never cause or initiate action for all overt intentional action.¹⁰

3 Libet Studies and the Proposed Implications for Free Action

Several philosophers and scientists, Libet included, have applied Libet et al.'s (1983) conclusions about the role of conscious intentions in action production to argue for free will skepticism. Free will skepticism is the view that there are no free agents; no agents met the

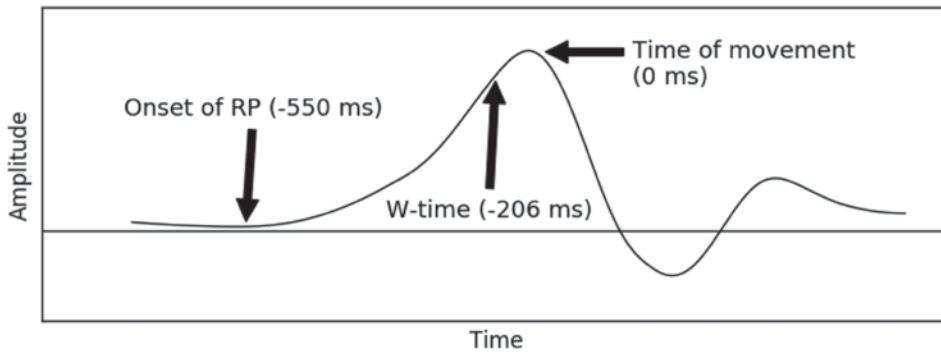


Figure 17.1 The Readiness Potential.

criteria to act freely. In this context, the thought is that if our conscious intentions never cause our actions, then we lack the control required to act freely and so to be blamed (for our bad decisions and actions) or praised (for our good decisions and actions). The phenomenology of agency might suggest that our conscious intentions cause our actions and hence seem to support that we decide and act freely at least some of the time, but our experience is mistaken. For instance, Libet has argued that there is no free will, only “free won’t.” That is, no one is morally accountable *for their intentions* to perform overt actions, as such intentions are formed prior to awareness, but agents can override their unconsciously acquired intentions and hence prevent such plans from being manifest (Libet 2005). If agents possess this negative capacity then individuals can still be held responsible for those overt bodily actions (1999, p. 54).¹¹

To grasp Libet’s claims about free will fully and the purported ways in which his results support free will skepticism more broadly, one must situate them in the particular account of free will that Libet endorses. First, a bit in the way of background: In the philosophical debate on free will, one major division is among compatibilists and incompatibilists about free will and determinism.¹² Determinism is the thesis that a complete, true description of the laws of nature and the state of the world at a time entails all other truths about that world. Compatibilists hold that free will is compatible with the truth of determinism.¹³ In contrast, incompatibilism is the thesis that the truth of determinism is incompatible with free will. That is, if determinism is true, then no one ever acts freely.¹⁴ In places, Libet seems committed to incompatibilism and to the further requirement of dualism of mental and physical properties (and perhaps substances) for free decision and action. For instance, he argues that if conscious intentions were efficacious in action production, then these intentions should “command the brain to perform the intended act” (Libet 2005, p. 553). Elsewhere he notes the following:

Are freely voluntary acts subject to macro-deterministic laws or can they appear without such constraints, nondetermined by natural laws and “truly free”? (2005, p. 551)

I myself proposed an experimental design that could test whether conscious will could influence nerve cell activities in the brain, doing so via a putative “conscious mental field” that could act without any neural connections as the mediators. (Libet 1994, 1999, p. 56)

Drawing upon the context of Libet's views on free will, there are at least three ways in which one could claim that Libet and colleagues' results threaten intentional action and free action.¹⁵ One could argue that the results support that the initiator of action, decisions to act, are neural events, but that intentional action and free action require a non-physical event of conscious decision among the initiators of action. Let's call this the DUALIST THREAT. Libet hints at such an argument in the first and last quote, and other neuroscientists have made similar claims, but this argument won't be assessed further here.¹⁶ Independently, one could argue that Libet et al. (1983) show that conscious decisions – regardless of whether such items are neural or of a non-physical nature – are epiphenomenal with respect to subsequent action, but that intentional and free action requires conscious intentions to be among the causes of intentional and free action. By epiphenomenal here, I mean that the conscious intention to *A* is caused by the neural events but is causally inefficacious (at least in relation to action *A*). Let's call this the EPIPHENOMENALISM THREAT. For instance, some neuroscientists have claimed that, despite the felt experience that our conscious intentions cause our overt actions, a conscious intention to *A* is just a post hoc construction on the basis of our already performed action of *A*-ing. I will address two studies cited in support of this argument and a further empirical study to the contrary in a later section. Given Libet's "free won't" proposal, this does not appear to be his view. Finally, one might argue on the basis of Libet's results, as Libet appears to, that our conscious intentions don't initiate the action production process (even if they are causally efficacious downstream in vetoing preparation for action) and that conscious intentions, not unconscious activity, must initiate action for an action to be free. Let's call this the UNCONSCIOUS INITIATION THREAT. This threat hinges crucially on the RP representing neural preparation for a committed particular action. Just what an RP signifies is challenged in empirical studies outlined below. Here the Epiphenomenalism Threat and the Unconscious Initiation Threat that the Libet paradigm poses to the existence of intentional and free action will be assessed via consideration of recent studies in the Libet paradigm. I will then offer a serious shortcoming of the paradigm that blocks generalizations to claims about intentional and free action. I will conclude by sketching how the studies might be modified to address this shortcoming, allowing better cross talk of neuroscience of free will and philosophical work on free will.

4 Notable Replications of the Libet Studies

Libet's work has spurred a flurry of related studies, replicating and extending the basic finding that relevant brain activity precedes *W* time. The Libet paradigm is a thriving one. As such it would be difficult to canvass all or even most of the neuroscience of agency studies inspired by Libet here. Rather, I'll mention two prominent replications of the study using fMRI (Soon et al. 2008) and single cell recording (Fried, Mukamel, and Kreiman 2011). I'll then address some extensions of the Libet paradigm that either support or challenge aspects of the studies with import to claims regarding intentional and free action.

One boldly titled article "Unconscious Determinants of Free Decisions in the Human Brain" replicates the general phenomenon of action-related neural activity preceding the reported time of awareness of intention to act. Soon and collaborators (2008) asked participants to decide, at a time of their choosing, to either press a button located on the left

with their left hand or press a button located on the right with their right hand. Similar to the original Libet task, participants were to hold in mind and report the time at which they first became aware of their decision to press the left (or right) button. However, instead of a Libet clock, participants viewed a rapidly updating stream of letters on screen, reporting the letter shown at the time of their awareness of their intention. Participants were given a wide window to produce “freely paced,” self-initiated responses, with button presses occurring on average 21.6 seconds after the start of the trial. Using fMRI data of the participants’ brain activity during the trials, Soon and his collaborators found that activity in the prefrontal and parietal cortex up to 10 seconds prior to the button push reliably predicted (above 50% chance) whether participants pressed the left or right button. In comparison, *W* time, here participants’ reported time of awareness of their intention to press the left (or right) button, was about 1 second prior to the button press. Soon and company marshal these data to support “the operation of a network of high-level control areas that begin to prepare an upcoming decision long before it enters awareness” (2008, p. 543). Like Libet, they take the unconscious brain activity, here activity in the prefrontal and parietal cortex, as representing an unconscious decision (“specifically encodes how a subject is going to decide” (2008, p. 545)). Further, from the title, we might provocatively conclude that free actions are unconsciously determined. Whether Soon and co-authors are warranted in concluding that neural activity 10 seconds prior to movement *determines* movement has been discussed (e.g., Mele 2012) and won’t be rehearsed here. Rather, in a later section I’ll present here a more widescope objection to this and other broadly classified Libet studies.

More recently, Fried, Mukamel, and Kreiman (2011) have replicated Libet’s finding that brain activity in the SMA and pre-SMA precedes *W* time using single cell recordings. Participants were epilepsy patients with electrodes implanted based on clinical need. In other respects, the experimental protocol mirrored Libet et al. (1983) with the substitution of a button press instead of a finger or wrist flex. The *W* time across trials, similar to Libet et al.’s results, was approximately 200 ms (mean of 193 ms; SD of 261 ms) prior to movement. Fried and company aligned neural spike trains in individual trials to that trial’s *W* time and analyzed when changes in neural activity deviated from baseline. Strikingly, Fried and collaborators found that neurons in the SMA, pre-SMA, and anterior cingulate – most significantly in the SMA and pre-SMA – exhibit gradual changes in firing rates (compared to baseline firing) as early as 1.5 seconds prior to *W* time. Further, Fried and collaborators’ results go beyond Libet et al.’s finding that unconscious related brain activity preceded reported time of conscious intention. The results support a significant relationship between early cited unconscious brain activity and *W*: In individual trials the activity of an assembly of 256 SMA neurons was predictive of *W* with 80% accuracy well before (700 ms) participants were aware of their intention to press the button. Hence, Fried and company argue that the activity of these neural populations is related to both the conscious decision to press the button as well as the button press itself.

Fried and colleagues’ innovative replication and extension of Libet’s original study suggest that the measured early brain activity in these neuroscience of intentional action studies is reflective of both volition and initiation of movement. This conclusion is consistent with both the Epiphenomenalism Threat and the Unconscious Initiation Threat, that conscious intentions to *A* are causally inert with respect to actions *A* and/or that actions in these studies are unconsciously initiated. However, recent neuroscientific

studies of intentional action have cast doubt on the role of the early measured unconscious neural activity (particularly in the SMA and pre-SMA) as specific to action or conscious intention to act.

5 Assessment of the Unconscious Initiation Threat

Several experiments have challenged whether the RP is a sufficient cause of the emergence of W and of conscious intentions to act.¹⁷ A 1999 experiment by Haggard and Eimer supports that the timing of the RP is not correlated with W time, but that timing of the LRP is: LRP is the lateralized motor activity. When an agent makes a unilateral movement, the LRP represents the difference in electrical potential to the left and right of the vertex prior to that movement. Haggard and Eimer utilized Libet's methodology for the experiments with the substitution of a free choice of left- or right- handed button press and then divided W times into early and late epochs. The mean RP did not differ significantly for early and late W epochs, but the mean LRP did. Drawing on Haggard and Eimer (1999), Alexander and colleagues (2014) ran related studies to test the nature of the relationship between RP/LRP and W. Using the same procedure as Haggard and Eimer, Alexander and company replicated the finding that RP profiles do not look significantly different for early versus late W times. However, they also found no covariance of the LRP and W, in contrast to Haggard and Eimer (1999). Notably, though, the data across the literature as to whether RP is sufficient for W is not univocal. Jo et al. (2015) found that the RP covaries with W time for a particular subpopulation, committed meditators.¹⁸

The most intriguing of the data challenging the sufficiency of the RP for W is a study by Schlegel and colleagues using hypnosis (2015): Highly hypnotizable individuals underwent EEG preparation and then hypnotic induction. During the task participants in both conditions watched a series of videos with a stress ball, palm up, in each hand. Left and right indicators appeared periodically to the sides of the video clips. In the standard condition, participants were told to squeeze the left stress ball if the indicator appeared on the left and to squeeze the right stress ball if the indicator appeared on the right. In comparison, those highly susceptible to hypnosis were subjected during induction to a post-hypnotic suggestion detailing this same instruction. Only participants who did not report being aware of this suggestion were included. Post-hypnosis they were placed in the experimental set-up and told to watch the clips. To account for the strangeness of their squeezing the stress balls, participants were informed post-hypnosis that the calibration of the electrodes might cause movement. Shlegel and collaborators found no significant differences in both RP and LRP profiles for the two conditions. These findings question whether there is any connection between the RP and W.¹⁹

Hence, we have reason to suspect that unconscious neural activity in the SMA and pre-SMA, at least as measured by the RP, has a weak, if any, connection to conscious practical decisions, decisions about what to do. In fact, several studies question whether the RP's role in action production has to do with movement-specific preparation at all and isn't rather more task general activity. If the RP, recorded in the SMA and pre-SMA, has a more general cognitive function than the preparatory process qua movement, the Unconscious Initiation Threat which points to the priority and causal efficacy of neural activity in the SMA and

pre-SMA is undermined. Recall that Libet and others take the RP to be indicative of an unconscious decision to act – an unconscious event of practical intention formation that initiates action.²⁰

Alexander and colleagues (2014) provide strong evidence that the RP isn't representative of movement-specific activity but that the LRP may be: Alexander et al. employed a Go/No Go set-up to test whether the RP and LRP are detectable when participants engage in a non-active task. Participants completed four kinds of trials, one in which they made a movement in response to a cue, one in which they prepared to make a movement but cancelled preparation if cued, another in which they predicted but did not report the color of a presented stimulus, and a fourth in which they added or subtracted numbers but did not report answers. Alexander and company measured the contingent negative variation (CNV) and the lateralized CNV (L-CNV) in all trials. The CNV is the negative potential signal for trials in which participants are cued to perform a task (analogous to the RP for endogenously generated trials). The L-CNV is the cued-trial analogue of the LRP. The CNV was present for all types of trials, but the L-CNV only appeared for trials inclusive of movement. Insofar as we can take the CNV and L-CNV to be analogous to the RP and LRP, these results support that the LRP is movement-specific but that the RP may represent more general task preparation.

Recently, Schurger, Jacobo, and Dehaene (2012) have pushed this skepticism about the role of the RP further, proposing the Integration to Bound model of RP. This model holds that, roughly, random walk-type fluctuation leads to neural activity crossing the threshold for motor response (i.e., movement). On this model RP does not reflect any true preparation.²¹ I'll say more about this result and a study in response briefly below.

The case for the proposal that early neural activity represents unconscious decisions to move in Libet studies is further weakened with related fMRI studies on perceptual decisions. Recall that Soon and colleagues (2008) claimed that neural activity up to 10 seconds prior to movement encoded a decision to press a left (right) button with the left (right) hand. Contra this claim, fMRI data support that neural activity in the parietal cortex – the localized brain activity referenced by Soon et al. (2008) – predicts perceptual guesses about which object is shown in uncertain conditions (Bode, Bogler, and Haynes 2013) and further are implicated in non-movement complex cognitive decisions, such as decisions whether to add or subtract numbers (Soon et al. 2013). Hence, the evidence suggests that it is a mistake to identify activity in this region of the parietal cortex with decisions with movement-specific content. Rather, the relevant activity in the parietal cortex reflects more general task preparation. Taken together, recent studies suggest that activity in the SMA, pre-SMA, and the posterior parietal cortex are not movement-specific in function, thus undermining the claim that said brain activity in these regions unconsciously initiate action. Perhaps we can only speculate that such regions aid in general preparation for cognitive decisions more broadly.²²

6 Assessment of the Epiphenomenalism Threat

Even if the Unconscious Initiation Threat can be neutralized,²³ one might worry that the Epiphenomenalism Threat remains. All of the outlined studies are consistent with conscious intentions to act being causally inert with respect to action. Mele (2009a) has provided some

evidence to the contrary when it comes to distal intentions to perform some task (2009a, pp. 134–135). (Distal intentions are intentions to do something later.) Further, Waller (2012) proposed modified experiments that could be run to test whether conscious intention *qua* conscious are causally efficacious via tasks in which conscious proximal intentions are plausibly required for successful execution of the task. However, the Epiphenomenalism Threat appears to be untested with regard to proximal intentions.

We can point to studies, though, that purport to support or undermine a particular version of the Epiphenomenalism Threat. For instance, both Lau, Rogers, and Passingham (2007) and Banks and Isham (2009) claim to have supported that conscious proximal intentions to move do not cause action but are created post-hoc from post-action events. For brevity's sake, I'll only detail one of the studies here: Banks and Isham (2009) asked participants to press a button at a time of their choosing and report the time of their awareness of their intention to press the button. The button was housed in a box, so that the button was not visible. Participants were told that a tone would sound when the button had been fully pressed. In actuality, a computer randomly generated the tone 5, 20, 40, or 60 ms following the button press. Banks and Isham found that participants' W time systematically moved forward in time in relation to the tone. That is, the greater the delay of tone post button press, the closer to 0 (objective time of the button press) participants reported becoming aware of their intention to press the button. Banks and Isham conclude that "the intuitive model [of a conscious intention causing action] has it backwards; generation of responses is largely unconscious, and we infer the moment of decision from the perceived moment of action (Eagleman 2004)" (2009, p. 20). On this view, one way to explain our illusion that conscious intentions are causally efficacious in action production is to note that there is a close temporal association of (consciously) intending to A and A-ing which misleads agents to infer a causal relation.²⁴ As Lau, Rogers, and Passingham (2007) argue, "One strong demonstration for the case of illusory conscious control would be that our perceived temporal order of intentions and actions are, in fact, false. If intentions, in fact, arise after actions, they could not, in principle, be causing actions" (81). If perception of action and its consequences contribute in some way to conscious intentions to act, then those conscious intentions could not have caused the action, the reasoning goes. Hence, Lau, Rogers, and Passingham (2007) and Banks and Isham (2009) conclude that conscious intentions, at least in Libet studies, are post-hoc constructions and so may be epiphenomenal.

The claim that the conscious intention to act in Libet studies is constructed post-action and so is causally inert in action production has met both conceptual and empirical resistance. As to the former, Mele (2009a, 2013) has challenged this conclusion in Lau, Rogers, and Passingham (2007) and in Banks and Isham (2009) on the grounds that what is at issue in the study is the participants' *belief* as to the timing of the conscious intention to press the button and that this belief may be mistaken: the time of the conscious intention itself may well remain unaffected by the manipulation. Further, Walsh, Long, and Haggard (2015) have empirically tested the claim regarding the post-hoc nature of conscious intentions with a phantom limb patient. Walsh and collaborators compared the features and presence of RP and W in a group of healthy participants versus a phantom limb patient using a button press Libet study set-up. The RP and W profile for the patient's missing left hand were not significantly different from the RP and W profile of her right hand and that

of control participants. These results speak against the claim that W and conscious intentions in Libet studies are constructed post-hoc: If it were the case, then given that the patient cannot execute the button press, the conscious intention to press the button (and perhaps also the RP in light of the contestable claims outlined above about the close connection between RP and W) should be significantly different in the phantom limb condition; The patient has no action or post-action event from which to construct the reported time of conscious intention. Of course, one might reasonably dispute, given the belief constraints on intention,²⁵ whether a phantom limb patient who is aware of her impairment could genuinely intend to move a limb that isn't there. Still, if we accept that the study involves conscious practical intentions, or at least W, Walsh and colleagues' findings are a mark against the post-hoc nature of conscious intentions in Libet studies and, perhaps, more broadly.

7 Generalization Worries and Future Directions

Thus far, I've outlined Libet's seminal study of intentional action and some claims regarding intentional and free action made on the basis of the Libet paradigm. I've canvassed the wider Libet study literature to assess support for the claims, arguing that prominent versions of the Epiphenomenalism Threat and the Unconscious Initiation Threat are likely false given conceptual arguments and empirical findings.

Still, there's further reason to be skeptical of both the Epiphenomenalism Threat and the Unconscious Initiation Threat apart from recent empirical tests: All of the outlined neuroscientific studies of free will haven't really captured, i.e., operationalized, free, and morally appraisable action. Given the nature of the actions examined, the studies fail to address or generalize to intentional actions characteristic of free and morally appraisable conduct, namely, those actions that feature the weighing of moral reasons or require conscious intentions (Mele 2009a; Waller 2012). Recount the tasks required of Libet study participants. One might worry that a spontaneous wrist or finger flex is too simple or arbitrary of an action to generalize the results of the studies. Past replications and extensions of the original Libet study also suffer from problematic operationalizations of relevant practical decision and action – e.g., when to press a button (Lau, Rogers, and Passingham 2007; Banks and Isham 2009; Fried, Mukamel, Kreiman 2011; Jo et al. 2015), which button to press (Haggard and Eimer 1999; Soon et al. 2008; Alexander et al. 2014), whether or not to press a button (Brass and Haggard 2008), cued simple movement (Alexander et al. 2014; Schlegel et al. 2015). The simple and arbitrary nature of these decisions and actions makes results drawn from the studies open to alternative interpretations regarding the import for free and morally appraisable agency.

A few neuroscientists of agency have turned their attention to addressing non-arbitrary (choosing, not picking) actions in the laboratory. Recall that Schurger, Sitt, and Dehaene (2012) offer the Integration to Bound model for the RP characteristic of intentional action in Libet-like studies, which suggests that the RP reflects mere stochastic fluctuation and not preparation for action (as was previously assumed). In response, Khalighinejad et al. (2018) adapted a perceptual decision-making paradigm similar to one used in primate studies (e.g., Shadlen and Roskies 2012), to test models of the RP as a true “preparatory process” against the Integration to Bound model. The experimental design includes a reward-incentivized

condition to the perceptual decision task, in which participants can choose to skip ahead to the start of the next trial for a small penalty. This experimental design thus notably sharpens the operational definition of volition from not just endogenously generated decisions and actions, but to ones made in light of reasons for action, with clear alternatives. Khalighinejad and company (2018) argue that the RP represents true preparation for movement given their results. Interestingly, Maoz et al. (2019), using a charity preference and donation paradigm similar to the study proposal in Waller (2012), found that trials involving deliberative decisions about moral matters evidence RPs with a significantly different amplitude or indeed no RPs at all when compared to trials involving arbitrary decisions. The LRP profiles on deliberative trials did not differ significantly from those on arbitrary decision trials. Participants were instructed to push one of two buttons to indicate their choice but were not asked to report the timing of their awareness of their decision in either kind of trial (Maoz et al. 2019). The results, then, are additional support for the movement-specific nature of the LRP and not the RP. Further, this study exhibits the rare merit of testing for morally weighty decisions that plausibly require deliberation.²⁶

Even given this marked improvement in methodology and sharpened conceptual distinctions – especially with regard to the kind of volition we care about – we still ought to be wary of any denials of free will and moral responsibility that rely on such results. As noted, the vast majority of studies of agency in the Libet paradigm do not concern the kinds of intentional actions for which agents are held accountable. The studies mentioned in this section offer more promising paradigms. However, the kinds of decisions and actions tested in Khalighinejad et al. (2018), while moving beyond arbitrary and simple actions to incentivized ones, are still not reflective of morally salient options.²⁷ Second, philosophers of action make an important distinction between practical decision-making – decisions about what to do – and theoretical (e.g., perceptual) decision-making – decisions about what is the case. The former is relevant to the volitional capacities at issue in discussions of free will and moral responsibility. If scientists are to theorize about human intentional action and free action, complex practical decision paradigms reflective of relevant sophisticated volitional capacities ought to be favored, when feasible, over perceptual decision-making paradigms. Third, participants' decisions should “make a difference” – that is, neuroscience of free will studies should incorporate decisions with actual meaningful consequences, not merely hypothetical ones.²⁸ As such, laboratory tests of the relevant kind of volition ought to stick as closely as possible to practical decision-making, of the conscious, morally valenced kind, inclusive of actual incentives, to maintain ecological validity. Maoz et al. (2019) are a step in this direction. Given the accumulating evidence that the RP is not movement-specific, future studies would do well to focus on the LRP and other neural markers of practical decision-making (e.g., using fMRI), especially in moral contexts or those in which conscious decisions are required for action execution. Such efforts would better support, then, generalizations to free and morally appraisable decisions and intentional actions.

Acknowledgments

I'd like to thank Al Mele, Kristin Michelson, Joseph Campbell, and participants at the Gothenburg Neuroscience and Responsibility Workshop, the London Mind Group, the King's College London Naturalised Accounts of Mind Group for their helpful feedback.

Notes

- 1 Non-human animals engage in goal-directed behaviors in some sense, but here the focus is a subset of goal-directed behaviors, human intentional actions.
- 2 For a good overview of the analysis of intentional action, see Mele (2009b).
- 3 For a good overview of the literature on intention, see Mele (2009b).
- 4 For explicit philosophical and psychological expressions of this commitment to the necessity of some conscious aspects of mental life for free will, see Mele (2010) and Baumeister (2008).
- 5 For representative accounts of free will that focus on decisions and actions informed by deliberation, see, among many, Fischer and Ravizza (1998), Kane (1996), and Mele (2006).
- 6 Libet's studies concern proximal intentions to act, intentions to act now. Proximal intentions can be distinguished from distal intentions to act, intentions to act sometime in the future (Mele 2009a, p. 10). A similar distinction is also made by Bratman (1984) (present-directed intentions and future directed intentions) and Pacherie (2006) (P-intentions and F-intentions).
- 7 Libet and his colleagues actually referred to this item in their instructions to participants as the broader disjunction "intention," "urge," or "desire" to flex (1983, p. 627), but I will refer to this item as an intention for continuity with the rest of the paper.
- 8 At least for the purposes of this paragraph. The arbitrary nature of these actions and their connection to free will will be assessed below.
- 9 I assess whether these critical assumptions are warranted in Waller (2019), Waller and Brager (2022), and Waller (2021).
- 10 Libet reasoned that if a RP is present for all overt movements and W appears after the onset of the RP in his studies, then RPs occur prior to conscious intentions and initiate action for all intentional actions, even complex ones (Libet 2005, p. 560).
- 11 For an critical empirical study of Libet's "free won't" or veto power, see Hallett (2011).
- 12 The compatibility question concerns whether free will is compatible with determinism. In the current context I understand this question to be one of the metaphysical relation between the relation of determinism and free will. I will remain open here as to whether the compatibilist and incompatibilist positions are ones of metaphysical incompatibility or metaphysical impossibility; the current debate, concerning the proposed condition that free will requires that conscious intentions to act cause overt intentional action, is plausibly a requirement on both compatibilist and incompatibilist accounts of free will. For an in-depth discussion of the relation between determinism and free will at issue in the broader literature, see Mickelson (2023).
- 13 It is important to note that most compatibilists do not think that the truth of determinism is required for agents to act freely, but merely that there is at least one possible world in which determinism is true of that world and some agent acts freely in that world.
- 14 Some free will theorists are agnostic about the compatibility question – are committed to neither compatibilism nor incompatibilism (e.g., Mele 1995).
- 15 Libet himself softens his stance in places, arguing for an instrumental adoption of the potential illusion of free will:

My conclusion that free will, one genuinely free in the non-determined sense, is then that its existence is at least as good, if not a better, scientific opinion than is its denial by determinist theory. Given the speculative nature of both determinist and nondeterminist theories, why not adopt the view that we have free will (until some real contradictory evidence may appear, if it ever does). Such a view would at least allow us to proceed in a way that accepts and accommodates our own deep feeling that we do have free will. (1999, p. 56).

- 16 An objection to arguments of this kind is taken up in Mele (2014).
- 17 In his critical discussion of the philosophical import of the Libet paradigm for free will, Mele helpfully distinguishes the time of the onset of the proximal intention to act, the time of the onset of the con-

- scious proximal intention to act, and the reported time of the conscious intention to act (W time) (2009a, p. 118). These times may or may not occur simultaneously in Libet studies, but experimenters often assume at least the latter two do without much support.
- 18 Committed meditators, the experimental group in Jo et al. (2015), were those who had at least three years of experience practicing mindfulness meditation. Other subpopulations show variance in W profile: Tabu and colleagues (2015) found that Parkinson's patients have delayed W time, but not delayed subjective reports of time of movement, compared to neurotypical controls.
 - 19 One might extend these results to question whether there are any conscious proximal intentions to flex or indeed any proximal intentions to flex full stop in this study. Mele (2009a) and Waller (2012) have questioned whether there are conscious proximal intentions in certain Libet paradigm studies. Both Mele (2009a) and Waller (2012) likewise question whether, assuming there are conscious proximal decisions in the relevant studies, the conscious aspect of these intentions is causally efficacious in producing action. Notably, Herdova (2016) argues that we have reason to be skeptical about the presence of any proximal intentions in Libet studies, conscious or not.
 - 20 Miller and Schwarz offer a different empirically motivated account of the RP, rejecting that the decision to act is made unconsciously in Libet studies. Rather, they argue for a spectrum of conscious awareness in which the RP represents "subcriterion levels of awareness rather than complete absence of awareness" (2014, p. 12).
 - 21 See also Murakami et al. (2014).
 - 22 This role is consistent with claims that the RP may reflect an urge to act or preparation for an urge to act, not a decision to act (see, e.g., Mele 2009a; Shepherd 2015).
 - 23 Of course, neuroscientists of agency and philosophers alike may well be able to argue that the activity of some other brain region R is implicated in preparation or in initiation of action prior to other awareness of a decision to act. Here we have only assessed this claim in relation to activity in the SMA, pre-SMA, and posterior parietal cortex.
 - 24 See also Wegner (2002).
 - 25 See Mele (2014) for an overview of some of the proposed belief constraints on intention.
 - 26 Of note, Pletti et al. (2015) did find RPs for deliberative decisions, albeit hypothetical decisions. Pletti and colleagues asked participants to make decisions about Trolley-like moral dilemmas, with instructions to either focus on legal considerations (Legal condition) versus no explicit instructions regarding how to deliberate (Non-legal condition). RP amplitude was lower for the Non-legal versus legal trials. The authors speculate that these results are due to the moral conflict causing a "reduced intention to act."
 - 27 Khalighinejad et al. (2018) don't make any claims about free will, but others may utilize their results in support of such claims. Marques (2015) makes a related point about Bode et al. (2011), arguing that such decisions are both passively made and don't require reflection on the part of the agent.
 - 28 See Harrison (2014) for the importance of paradigms with actual choices.

Bibliography

- Alexander, P., Schlegel, A., Sinnott-Armstrong, W., Roskies, A., Tse, P., and Wheatley, T. (2014). Dissecting the readiness potential. In: *Surrounding Free Will: Philosophy, Psychology, Neuroscience* (ed. A.R. Mele), 203–230. New York: Oxford University Press.
- Banks, W.P. and Isham, E.A. (2009). We infer rather than perceive the moment we decided to act. *Psychological Science* 20: 17–21. doi: 10.1111/j.1467-9280.2008.02254.x.
- Baumeister, R. (2008). Free will, consciousness, and cultural animals. In: *Are We Free? Psychology and Free Will* (ed. J. Baer, J.C. Kaufman and R. Baumeister), 65–85. New York: Oxford University Press.

- Bode, S., Bogler, C., and Haynes, J.-D. (2013). Similar neural mechanisms for perceptual guesses and free decisions. *Neuroimage* 65: 456–465. doi: 10.1016/j.neuroimage.2012.09.064.
- Bode, S., He, A.H., Soon, C.S., Trampel, R., Turner, R., and Haynes, J.D. (2011). Tracking the unconscious generation of free decisions using ultra-high field fMRI. *PloS one* 6 (6): e21612.
- Brass, M., and Haggard, P. (2008). The what, when, whether model of intentional action. *The Neuroscientist* 14 (4): 319–325.
- Bratman, M.E. (1984). Two faces of intention. In: *Oxford Readings in Philosophy: The Philosophy of Action* (ed. A.R. Mele), 178–203. Oxford: Oxford University Press.
- Eagleman, D.M. (2004). The when and where of intention. *Science* 303: 1144–1146. doi: 10.1126/science.1095331.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.
- Fischer, J.M. and Tognazzini, N. (2011). The physiognomy of responsibility. *Philosophy and Phenomenological Research* 82 (2): 381–417. doi: 10.1111/j.1933-1592.2010.00458.x.
- Fried, I., Mukamel, R., and Kreiman, G. (2011). Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron* 69 (3): 548–562. doi: 10.1016/j.neuron.2010.11.045.
- Haggard, P. and Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research* 126 (1): 128–133.
- Hallett, M. (2011). Volition: how physiology speaks to the issue of responsibility. In: *Conscious Will and Responsibility* (ed. W. Sinnott-Armstrong and L. Nadel), 61–69. Oxford: Oxford University Press.
- Harrison, G. (2014). Real choices versus hypothetical choices. In: *Handbook of Choice Modelling* (ed. S. Hess and A. Daly), 236–256. Cheltenham, UK: Edward Elgar Publishing.
- Herdova, M. (2016). Are intentions in tension with timing experiments? *Philosophical Studies* 173 (3): 573–587. doi: 10.1007/s11098-015-0507-0.
- Jo, H.-G., Hinterberger, T., Wittmann, M., and Schmidt, S. (2015). Do meditators have higher awareness of their intentions to act? *Cortex* 65: 149–158. doi: 10.1016/j.cortex.2014.12.015.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.
- Khalighinejad, N., Schurger, A., Desantis, A., Zmigrod, L., and Haggard, P. (2018). Human self-initiated action is preceded by a reliable process of noise reduction. *Neuroimage* 165: 35–47. doi: 10.1016/j.neuroimage.2018.01.015.
- Kornhuber, H.-H. and Deecke, L. (1965). Changes in the brain potential in voluntary movements and passive movements in man: readiness potential and reafferent potentials. *Pflügers Archiv Für Die Gesamte Physiologie Des Menschen Und Der Tiere* 284: 1–17.
- Lau, H.C., Rogers, R.D., and Passingham, R.E. (2007). Manipulating the experienced onset of intention after action execution. *Journal of Cognitive Neuroscience* 19: 81–90. doi: 10.1162/jocn.2007.19.1.81.
- Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences* 8: 529–566.
- Libet, B. (1994). A testable field theory of mind-brain interaction. *Journal of Consciousness Studies* 1: 119–126.
- Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies* 6 (8–9): 47–57.
- Libet, B. (2005). Do we have free will? In: *The Oxford Handbook of Free Will* (ed. R. Kane), 551–564. Oxford: Oxford University Press.
- Libet, B. (2011). Do we have free will? In: *Conscious Will and Responsibility*, (eds. W. Sinnott-Armstrong and L. Nadel), 1–10. Oxford: Oxford University Press.
- Libet, B., Gleason, C.A., Wright, E.W., and Pearl, D.K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). *Brain* 106 (3): 623–642.
- Libet, B., Wright, E.W., and Gleason, C.A. (1982). Readiness potentials preceding unrestricted spontaneous pre-planned voluntary acts. *Electroencephalography and Clinical Neurophysiology* 54: 322–335.

- Maoz, U., Yaffe, G., Koch, C., and Mudrik, L. (2019). Neural precursors of decisions that matter—an ERP study of deliberate and arbitrary choice. *ELife* 8: e39787. doi: 10.1101/097626.
- Marques, B.S. (2015). Different kinds of decisions and an experiment on unconscious generation of free decisions: a conceptual analysis. *Filosofia Unisinos* 16 (1): 44–57. doi: 10.4013/fsu.2015.161.03.
- Mele, A. (2014). Free will and substance dualism: the real scientific threat to free will? In: *Moral Psychology: Free Will and Moral Responsibility* (ed. W. Sinnott-Armstrong), 195–208. Cambridge, MA: MIT Press.
- Mele, A.R. (1995). *Autonomous Agents: From Self-Control to Autonomy*. Oxford: Oxford University Press.
- Mele, A.R. (2006). *Free Will and Luck*. New York: Oxford University Press.
- Mele, A.R. (2009a). *Effective Intentions*. Oxford: Oxford University Press.
- Mele, A.R. (2009b). Intention and intentional action. In: *Oxford Handbook of Philosophy of Mind* (ed. B.P. McLaughlin, A. Beckermann and S. Walter), 691–712. Oxford: Oxford University Press.
- Mele, A.R. (2010). Testing free will. *Neuroethics* 3: 161–172.
- Mele, A.R. (2012). Another scientific threat to free will? *The Monist* 95 (3): 422–440. doi: 10.5840/monist201295322.
- Mele, A.R. (2013). Vetoing and consciousness. In: *Decomposing the Will* (ed. A. Clark, J. Kiverstein and T. Vierkant), 73–86. Oxford: Oxford University Press.
- Miller, J. and Schwarz, W. (2014). Brain signals do not demonstrate unconscious decision making: an interpretation based on graded conscious awareness. *Consciousness and Cognition* 24: 12–21. doi: 10.1016/j.concog.2013.12.004.
- Mickelson, K. (2023). (In)compatibilism. In: *A Companion to Free Will* (eds. Campbell, J.K., Mickelson, K.M. and White, V.A.), 1–29. Hoboken, NJ, USA: Wiley Blackwell.
- Murakami, M., Vicente, M.I., Costa, G.M., and Mainen, Z.F. (2014). Neural antecedents of self-initiated actions in secondary motor cortex. *Nature Neuroscience* 17 (11): 1574–1582. doi: 10.1038/nn.3826.
- Pacherie, E. (2006). Towards a dynamic theory of intentions. In: *Does Consciousness Cause Behavior? An Investigation of the Nature of Volition* (ed. S. Pockett, W.P. Banks and S. Gallagher), 145–167. Cambridge, MA: MIT Press.
- Pletti, C., Sarlo, M., Palomba, D., Rumiati, R., and Lotto, L. (2015). Evaluation of the legal consequences of action affects neural activity and emotional experience during the resolution of moral dilemmas. *Brain and Cognition* 94: 24–31. doi: 10.1016/j.bandc.2015.01.004.
- Schlegel, A., Alexander, P., Sinnott-Armstrong, W., Roskies, A., Tse, P.U., and Wheatley, T. (2015). Hypnotizing libet: readiness potentials with non-conscious volition. *Consciousness and Cognition* 33: 196–203. doi: 10.1016/j.concog.2015.01.002.
- Schurger, A., Sitt, J.D., and Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences* 109 (42): E2904–E2913. doi: 10.1073/pnas.1210467109.
- Shadlen, M.N. and Roskies, A.L. (2012). The neurobiology of decision-making and responsibility: reconciling mechanism and mindedness. *Frontiers in Neuroscience* 56: 1–12. doi: 10.3389/fnins.2012.00056.
- Shepherd, J. (2015). Scientific challenges to free will and moral responsibility. *Philosophy Compass* 10 (3): 197–207. doi: 10.1111/phc3.12200.
- Soon, C.S., Brass, M., Heinze, H.-J., and Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience* 11: 543–545.
- Soon, C.S., He, A.H., Bode, S., and Haynes, J.-D. (2013). Predicting free choices for abstract intentions. *Proceedings of the National Academy of Sciences* 110 (15): 6217–6222. doi: 10.1073/pnas.1212218110.
- Tabu, H., Aso, T., Matsushashi, M., Ueki, Y., Takahashi, R., Fukuyama, H., Shibasaki, H., and Mima, T. (2015). Parkinson’s disease patients showed delayed awareness of motor intention. *Neuroscience Research* 95: 74–77. doi: 10.1016/j.neures.2015.01.012.
- Waller, R.R. (2012). Beyond button presses: the neuroscience of free and morally appraisable actions. *The Monist* 95 (3): 441–462. doi: 10.5840/monist201295323.

- Waller, R.R. (2019). Recent work on agency, freedom, and responsibility: a review. John Templeton Foundation. <https://www.templeton.org/wp-content/uploads/2019/10/Free-Will-White-Paper.pdf>. Accessed on 19th October 2020.
- Waller, R.R. (2021). Weighing in on decisions in the brain: neural representations of pre-awareness practical intention. *Synthese* 199 (1): 5175–5203.
- Waller, R.R. and Brager, A.J. (2022). I did that!: biomarkers of volitional and free agency. In: *Neuroscience and Philosophy* (ed. W. Sinnott-Armstrong and F. De Brigard). Cambridge, MA: MIT Press.
- Walsh, E., Long, C., and Haggard, P. (2015). Voluntary control of a phantom limb. *Neuropsychologia* 75: 341–348. doi: 10.1016/j.neuropsychologia.2015.06.032.
- Watson, G. (1996). Two faces of responsibility. *Philosophical Topics* 24: 227–248.
- Wegner, D.M. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

A Defense of Natural Compatibilism

FLORIAN COVA

1 Investigating Folk Conceptions of Free Will and Moral Responsibility

If we define FREE WILL as the type of control one has to exert upon one's action to be morally responsible for them (see for example Mele 2008),¹ then FREE WILL is not a technical concept coined by philosophers, such as ALIEFS, EMERGENCE, or TROPES. Rather, such a concept seems to be already present in our everyday practices – at least at an implicit level, where it guides our assessment of others' moral responsibility, the reactive attitudes we adopt towards them and certain behaviors such as rewards and punishment. Long before Aquinas introduced the expression *liberum arbitrium*, people were distinguishing between people who acted willfully and those who acted under duress, and responded to their actions accordingly.

Thus, it is not a stretch to say that philosophical reflection on free will takes its roots in an intuitive notion that pervades our interactions with others. Recently, philosophers and psychologists have tried to get a better understanding of this everyday understanding of free will. Their research has focused on two different questions. The first one is *what kind of control do people think we need to have over our actions to be morally responsible for them?* The second is *what kind of control do people think we actually have on our actions?* In this chapter, I will focus on the first question, as it is the one that has generated the most research and debate in the past years (though I will briefly touch on the second towards the end of the chapter).

1.1 The Relevance of Folk Intuitions to the Philosophical Debate on Free Will

Besides the obvious psychological importance of these questions for our understanding of the human mind and social cognition, one might wonder what is the relevance of such inquiries for philosophical theorizing about free will. An answer is that intuitions about the conditions required for moral responsibility play a major role in philosophical debates about the *nature* of free will: most arguments take as their premise either principles that are supposed to be self-evident (e.g., the Principle of Alternate Possibilities, van Inwagen's Rule Beta) or intuitions about individual thought-experiments (e.g., Frankfurt cases, Manipulation arguments).

A Companion to Free Will, First Edition. Edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White.
© 2023 John Wiley & Sons, Inc. Published 2023 by John Wiley & Sons, Inc.

However, despite some attempts (e.g., Double 1996), the free will debate is still in need of an explicitly articulated meta-philosophy, and it is not clear what the status and role of such intuitions are. There are several ways to understand the role intuitions play in the philosophical debate about the nature of free will. According to what I call the *weak view*, intuitions only set the default point and determine on who lies the burden of proof (e.g., Nahmias et al. 2006). According to the *moderate view*, intuitions give us access to certain modal truths about free will and moral responsibility (e.g., that it is *possible* for someone to be morally responsible even if one could not have done otherwise) and constitute *prima facie* defeasible evidence in their favor. Finally, according to what I call the *strong view*, intuitions have the power to determine the subject matter of the free will debate, either because widespread intuitions and truisms about free will and moral responsibility actually determine the reference of the expression “free will,” or because these intuitions provide the main basis for a conceptual analysis of our shared concept of FREE WILL, in a way that makes it impossible for an adequate theory of free will to ignore our intuitions about it (e.g., Jackson 1998). Whatever your position on these matters, it seems clear that getting a better grasp of the intuitive ground from which the philosophical problem of free will emerged might prove useful in evaluating certain philosophical arguments.

1.2 Two Main Positions: Natural Compatibilism vs. Natural Incompatibilism

In their study of folk intuitions about the *nature* of free will, experimental philosophers and psychologists have mainly focused on the *compatibility question*: do people consider determinism to prevent the possibility of free will? *Natural Incompatibilism* is the claim that laypeople conceive free will in such a way that it is incompatible with determinism, because they take determinism to be an obstacle to free will. *Natural Compatibilism* is the claim that free will as we intuitively think of it is perfectly compatible with determinism. Though some have argued that the very same people can sometimes have compatibilist and incompatibilist intuitions depending on the context (Doris, Knobe and Woolfolk 2007; Knobe and Nichols 2010; Cova 2011), most of the debate in the literature has been framed as an opposition between *Natural Compatibilism* and *Natural Incompatibilism*. In this chapter, I will review this experimental literature and argue that, if we were to choose between the two, *Natural Compatibilism* would be a better fit of the available data.

But before we start, I need make two additional remarks. The first one is about the material I will cover in this chapter. As pointed out by Cova and Kitano (2014) and Feltz (2016), these studies come in two kinds. A first group set of studies tries to *directly* address the compatibility question, by investigating whether people think that an agent in a deterministic universe can be free and morally responsible. A second group *indirectly* addresses the question by studying whether people share the intuitions that serve as premise to philosophical arguments for and against the compatibility of free will with determinism, such as Frankfurt-style cases (Miller and Feltz 2011; Cova 2014, 2017) or Manipulation arguments (Sripada 2012; Feltz 2013; Björnsson 2016; Cova forthcoming). In this chapter, I focus on studies that try to *directly* address the compatibility question.

The second remark is about the way these studies have operationalized and measured free will. Some studies directly ask participants about free will (e.g., by asking them whether the agent *freely* did a certain thing, or did a certain thing *of his own free will*), other ask participants about moral responsibility (e.g., by asking whether the agent is *morally responsible* for doing something, or deserves *blame/praise* or a *punishment/reward* for what they did), and others ask both types of questions. As mentioned at the start of this chapter, I will focus on free will defined as the type of control required by moral responsibility – meaning that,

when both kinds of measure will yield different conclusions, I will take participants' answers about moral responsibility as the most reliable indicator. Indeed, though investigating how participants use terms such as "free will" or "freely" might be interesting in its own right, it is not clear how these are relevant to the philosophical debate, as very few philosophical arguments seem to depend on linguistic intuitions about the way such words are used in everyday language. Moreover, while practices involving attributions of moral responsibility (such as reactive attitudes or punishment) seem pervasive and human universals, some cultures seem to lack words or expressions that would be equivalent to "free will" (Berniūnas et al. 2021). Finally, most people take philosophical debates about free will to be important precisely because of their practical relevance and the implications they have for our image of ourselves as morally responsible agents. For all these reasons, I will thus focus on participants' intuitions about moral responsibility when possible.

2 A Short History of the Natural Compatibilism vs. Natural Incompatibilism Debate

2.1 *Folks as Natural Compatibilists*

The first studies on the compatibility question were published in 2005 and 2006 by Nahmias and his colleagues (Nahmias et al. 2005, 2006). The principle of these studies was quite simple: participants were presented with a description of a deterministic universe, then were told about an agent living in this universe performing a particular action. Participants were then asked whether the agent was morally responsible for this action, and acted of his own free will.

Here is an example:

SUPERCOMPUTER – Imagine that in the next century we discover all the laws of nature, and we build a supercomputer which can deduce from these laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time. It can look at everything about the way the world is and predict everything about how it will be with 100% accuracy. Suppose that such a supercomputer existed, and it looks at the state of the universe at a certain time on March 25, 2150 AD, 20 years before Jeremy Hall is born. The computer then deduces from this information and the laws of nature that Jeremy will definitely rob Fidelity Bank at 6:00 pm on January 26, 2195. As always, the supercomputer's prediction is correct; Jeremy robs Fidelity Bank at 6:00 pm on January 26, 2195.

Some participants were then asked:

Imagine such a supercomputer actually did exist and actually could predict the future, including Jeremy's robbing the bank (and assume Jeremy does not know about the prediction):
Do you think that, when Jeremy robs the bank, he acts of his own free will?

While others were asked:

Do you think that when Jeremy robs the bank, he's morally blameworthy for it?

To the first question, 76% of participants answered that Jeremy acted of his own free will. To the second, 83% of participants answered that Jeremy was morally blameworthy. In another version, Jeremy did not rob a bank but saved a child from a burning building. To this scenario,

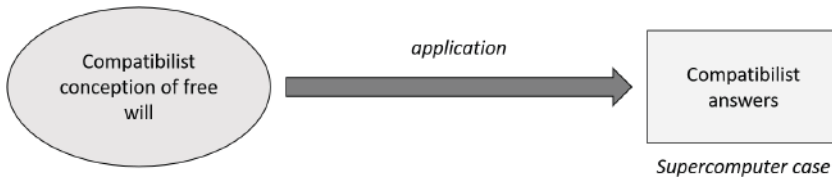


Figure 18.1 Folks as natural compatibilists.

68% of participants answered that Jeremy saved the child of his own free will and 88% judged that Jeremy was praiseworthy for having saved the child. Finally, in a third version of the story, Jeremy decided to go jogging. In this case, 79% of participants answered that Jeremy went jogging of his own free will.

Nahmias and his colleagues obtained similar results for scenarios describing a universe in which one's actions and decisions are fully determined by the combination of one's genes and upbringing, and for scenarios describing a universe submitted to *Eternal Recurrence*, where the same things are doomed to happen again and again. Together, these results suggest that *most* (though not all) of their participants have compatibilist intuitions: they judge free will and moral responsibility to be compatible with determinism (see Figure 18.1).

2.2 Folks as Natural Incompatibilists: The Performance Error Model

However, things might not be so straightforward. Noticing that the scenarios used by Nahmias and his colleagues all make use of "concrete" cases, in which participants are asked about a single action performed by a given, identifiable individual, Nichols and Knobe (2007) investigated how participants would respond to more "abstract" questions. They first presented participants with the following description of two universes:

Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe caused what happened next, and so on right up until the present. For example, one day John decided to have French Fries at lunch. Like everything else, this decision was completely caused by what happened before it. So, if everything in this universe was exactly the same up until John made his decision, then it had to happen that John would decide to have French Fries.

Now imagine a universe (Universe B) in which almost everything that happens is completely caused by whatever happened before it. The one exception is human decision making. For example, one day Mary decided to have French Fries at lunch. Since a person's decision in this universe is not completely caused by what happened before it, even if everything in the universe was exactly the same up until Mary made her decision, it did not have to happen that Mary would decide to have French Fries. She could have decided to have something different.

The key difference, then, is that in Universe A every decision is completely caused by what happened before the decision – given the past, each decision has to happen the way that it does. By contrast, in Universe B, decisions are not completely caused by the past, and each human decision does not have to happen the way that it does.

Obviously, Universe A is intended to be a deterministic universe, while Universe B, by contract, is intended to be an indeterministic universe. After reading this description, participants were randomly assigned to the *concrete* or *abstract* condition. Participants in the *concrete* condition received the following vignette:

In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and 3 children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family.

Is Billy fully morally responsible for killing his wife and children?

In this condition, most participants (72%) gave the compatibilist answer according to which the agent was fully morally responsible, even if he lived in deterministic universe A. These results are consistent with those obtained by Nahmias and his colleagues. But, let's now consider the *abstract* condition. Participants in this condition had no scenario to read but just received the following question:

In Universe A, is it possible for a person to be morally responsible for their actions?

In this condition, most participants (86%) gave the *incompatibilist* answer. To put it otherwise: participants tended to deny that agents could be morally responsible for their action in the deterministic universe A. Thus, participants' answers to experimental philosophers' probes seem incoherent. How are we to explain these results?

Nichols and Knobe have their own explanation; the Performance Error Model. According to this model, the folk conception of free will is, at its core, perfectly incompatibilist: free will, as we naturally conceive it, is incompatible with determinism. This commitment is what is reflected in the abstract condition, when participants are not asked to judge and condemn a particular individual: they give mostly *incompatibilist* answers. However, when faced with a concrete case (that is: a case featuring a given agent performing a precise action), participants' intuitions and judgments can be swayed by their emotions. Indeed, particular actions can be revolting and outrageous (as Billy's murder of his wife and three children), thus eliciting the desire to punish the agent. This desire, in turn, leads people to justify their desire to punish the agent by attributing him a certain amount of moral responsibility (Figure 18.2).

Nichols and Knobe's Performance Error Model is in line with other findings in social psychology, namely that people are more likely to believe in free will when they have just

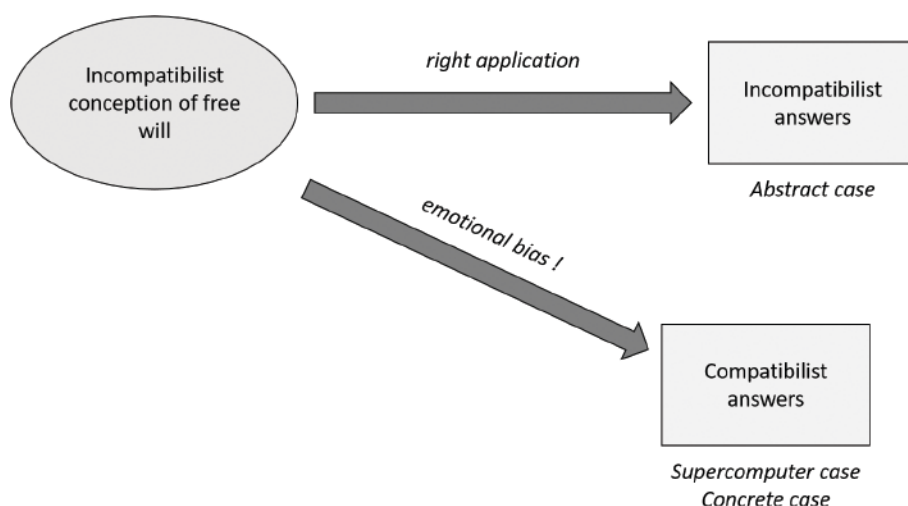


Figure 18.2 Nichols and Knobe's Performance Error Model.

been presented with descriptions of moral violations (Clark et al. 2014). It is also supported by additional evidence presented by Nichols and Knobe themselves. Indeed, one prediction of the Performance Error Model is that participants should give mostly *incompatibilist* answer when presented with a concrete case that is not emotionally salient. To test this prediction, Nichols and Knobe designed two new conditions. The *low affect* condition was the following:

As he has done many times in the past, Mark arranges to cheat on his taxes. Is it possible that Mark is fully morally responsible for cheating on his taxes?

While the *high affect* condition was the following:

As he has done many times in the past, Bill stalks and rapes a stranger. Is it possible that Bill is fully morally responsible for raping the stranger?

Participants were assigned either to the *low affect* or *high affect* condition, and told either that the agent lived in the (deterministic) Universe A or that he lived in the (indeterministic) Universe B. Table 18.1 describes, for each case, the proportion of participants who judged that the agent could be fully morally responsible for his action.

As predicted by Nichols and Knobe's Performance Error Model, participants in the *low affect* condition tended to judge that the agent could not be responsible for cheating on his taxes when he lived in a deterministic universe. On the contrary, participants in the *high affect* condition judged that the agent was morally responsible for raping his victim, even when he lived in a deterministic universe. Thus, it seems that participants' intuitions perfectly fit the Performance Error Model.

However, there is a problem with Nichols and Knobe's results: other experimental philosophers have had a lot of trouble replicating them. Indeed, though the difference between the *abstract* and *concrete* condition seems robust (and has been replicated in several countries, see Sarkissian et al. 2010), other experimental philosophers had trouble replicating the difference between the *low affect* and *high affect* conditions. A meta-analysis by Feltz and Cova (2014), which aggregated all known attempts at replicating this effect, found that the real difference between the *low* and *high* affect conditions was actually very small – and thus that the effect of emotional reactions on judgments about moral responsibility were not strong enough to actually explain the huge difference between the *abstract* and *concrete* condition.

Another problem for the Performance Error Model comes from a study conducted by Cova, Bertoux, and their colleagues (2012) on patients suffering from a behavioral variant of frontotemporal dementia, a neurodegenerative disease accompanied by a deficit in emotional responses. Given bvFTD patients' impaired emotional reactions to descriptions of gruesome and violent acts, the Performance Error Model should predict that bvFTD patients, being free of emotional biases, should give more incompatibilist answers than control participants to concrete cases. Thus, Cova and his colleagues presented participants with the *supercomputer* and *concrete* cases. However, bvFTD patients gave the same answers as control participants – that is: mostly compatibilist answers.

Table 18.1 Results from Nichols and Knobe (2007).

	Agent in deterministic universe A	Agent in indeterministic universe B
High Affect	64%	95%
Low Affect	23%	89%

Thus, it seems that participants' compatibilist answers cannot be explained away as the mere effect of emotional biases. We thus need another explanation for the apparent incoherence in participants' judgments about moral responsibility.

2.3 *Folks as Natural Compatibilists: The Bypassing Hypothesis*

This other explanation is Murray and Nahmias' Bypassing Hypothesis (Murray and Nahmias 2014). As the Performance Error Model, the Bypassing Hypothesis was originally designed to account for an apparent incoherence in participants' judgments. Indeed, as a follow-up on his prior investigations, Nahmias (2006) decided to investigate whether the *kind* of determinism participants were presented with would make a difference (see also Nahmias, Coates and Kvaran 2007). This is why he used a pair of scenarios describing two different kinds of determinism. The first scenario (*psychological determinism*) described a planet similar to ours (Ertan) inhabited by people called the Ertans. On this planet, Ertan psychologists have discovered that the Ertan's thoughts, desires, and plans occurring in her or his mind completely cause all the decision the Ertan makes. The psychologists also have discovered that these thoughts, desires, and plans are completely caused by the Ertan's current situation and the antecedent events she or he has been through. In the second scenario (*neurological determinism*), the same planet was described but, this time, the neuroscientists have discovered that the decision the Ertan makes is completely caused by the specific neural processes occurring in his or her brain; these neural processes are completely caused by the Ertan's current situation and the antecedent events she or he has been through.

For both scenarios, participants were asked if Ertans could act of their own free will and whether they deserved to be given credit or blame for their actions. In the *psychological determinism* condition, 72% of participants answered positively to the first question and 77% answered positively to the second question, thus giving perfectly compatibilist answers. In the *neurological determinism* condition, only 18% of participants answered positively to the first question and 19% to the second question, that is: clearly incompatibilist answers. To put it otherwise: participants gave mostly compatibilist answers in the first case and mostly incompatibilist answers in the second case. Once again, there seems to be something incoherent in participants' judgments about free will and moral responsibility.

However, Nahmias offers a plausible explanation for this pattern of judgment. According to him, this asymmetry arises because people take neurological determinism, but not psychological determinism, to imply that people's mental states play no role in the generation of their decisions and actions – that people's mental state are, so to speak, *bypassed*. But, both compatibilists and incompatibilists alike would agree that we would not be free nor morally responsible for our actions if our desires, belief, and values played no role in our decisions and actions. Thus, participants' seemingly "incompatibilist" answers to *neurological determinism* case do not show that they have an incompatibilist understanding of free will – only that (i) they take neurological determinism to imply "bypassing" and (ii) that they take "bypassing" to exclude free will and moral responsibility. On the other hand, their compatibilist answers to the *psychological determinism* case reflect their true, compatibilist understanding of free will.

According to Murray and Nahmias (2014, see also Nahmias and Murray 2011), the same kind of explanation can be applied for the difference between Nichols and Knobe's *abstract* and *concrete* conditions. For them, Nichols and Knobe's description of determinism is misleading and lead participants to assume not only that Universe A is deterministic (i.e., that every human action in it is fully caused by prior events), but also that it involves "bypassing" (i.e., that agents' mental states do not play a causal role in the generation of their actions and decisions). This is why participants who are presented with this case and asked the *abstract*

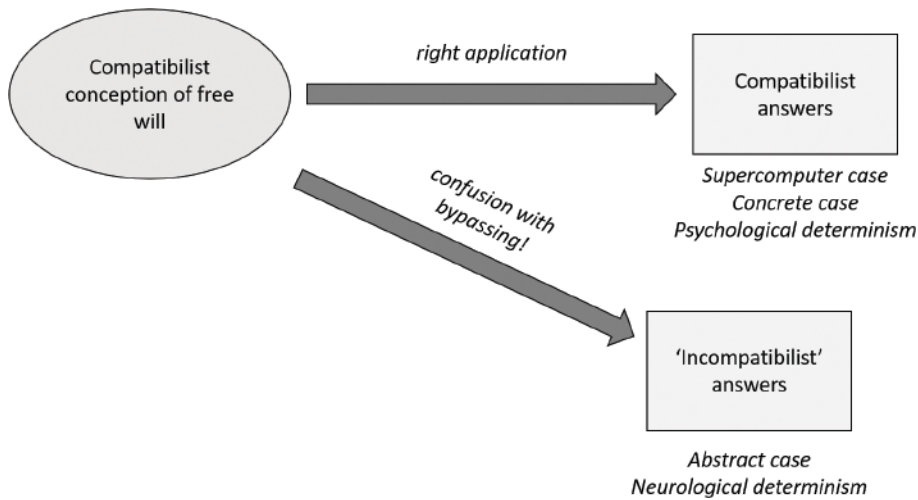


Figure 18.3 Nahmias and Murray's Bypassing Hypothesis.

question give seemingly “incompatibilist” answers: they are just expressing their belief that moral responsibility is incompatible with “bypassing.”

But why do participants tend to give compatibilist answer in the *concrete* case, then? Here, the idea is that, because the concrete case clearly states that the agent acts on the basis of his mental states difference (he kills his wife and children because he *wants* to be with his secretary), participants might be less likely to infer that determinism entails bypassing in these cases. Hence the mostly compatibilist answers (see Figure 18.3).

To test these predictions, Nahmias and Murray (2011) presented participants either with Nichols and Knobe's *abstract* and *concrete* condition (in Universe A), or with an abstract and a concrete version of the eternal recurrence case used by Nahmias and his colleagues (2006). For each scenario, participants were not only asked if agents in these scenarios deserved praise or blame and acted from their own free will, but they were also asked questions designed to probe their understanding of determinism, and to which extent they were likely to confuse determinism with bypassing:

Decisions: In Universe [A/C], a person's decisions have no effect on what they end up being caused to do.

Wants: In Universe [A/C], what a person wants has no effect on what they end up being caused to do.

Believes: In Universe [A/C], what a person believes has no effect on what they end up being caused to do.

No Control: In Universe [A/C], a person has no control over what they do.

Participants' answers to these questions were aggregated to constitute a *bypassing* score. The higher the score, the more participants were likely to confuse determinism with bypassing. Then, Nichols and Murray conducted a *mediation analysis* to determine to which extent participants' tendency to confuse determinism with bypassing explained the impact of different scenarios on participants' judgments about free will and moral responsibility. The results of this analysis (presented in Figure 18.4) revealed that bypassing scores *fully mediated* the effect of scenarios on judgments of free will and moral responsibility, suggesting that participants' incompatibilist judgments are indeed explained by their taking determinism to imply bypassing. To put it

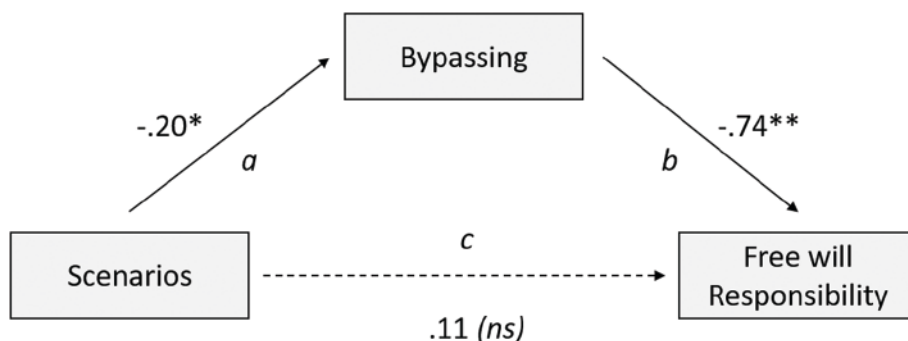


Figure 18.4 Mediation analysis by Nahmias and Murray (2011).

otherwise: seemingly incompatibilist judgments seem to be the product of a confusion, and when people understand exactly what determinism is, they tend to give compatibilist answers.

3 Answering Methodological Objections to the Bypassing Hypothesis

3.1 Two Objections to the Bypassing Hypothesis

The Bypassing Hypothesis seems to support the idea that people are mostly compatibilists at heart (what we called *Natural compatibilism*). However, for all its merits, the Bypassing Hypothesis has drawn several criticisms.

A first set of criticisms comes from Rose and Nichols (2013). They argue that Nahmias and Murray's model is not the only (nor the best) explanation for their results. Indeed, it turns out that, re-analyzing Nahmias and Murray's results, one can find that free will and moral responsibility scores (henceforth: FW/MR scores) also mediate the effect of scenarios on bypassing scores. So, Nahmias and Murray's data are compatible with two different models:

- The Bypassing Hypothesis: Scenarios $\rightarrow$ Bypassing scores $\rightarrow$ FW/MR scores
- The Incompatibilist Model: Scenarios $\rightarrow$ FW/MR scores $\rightarrow$ Bypassing scores

To determine which of these two models is the best, Rose and Nichols recruited participants that were presented with Nichols and Knobe's *abstract* condition, either situated in Universe A or Universe B. Participants were asked questions about free will, moral responsibility, and bypassing. Rose and Nichols then used structural equation modelling and the TETRAD IV program to compare the Bypassing Hypothesis and the Incompatibilist Model. They found that the Incompatibilist Model was a better fit of the data than the Bypassing Hypothesis (see Figure 18.5).

A second, related set of criticisms was put forward by Björnsson (2014). Björnsson ran a new study using questions about free will, moral responsibility and bypassing, but also introducing a new type of question, which he calls *throughpass questions*:

Abstract Throughpass: In Universe A, when earlier events cause an agent's action, they typically do so by affecting what the agent believes and wants, which in turn causes the agent to act in a certain way.

Concrete Throughpass: When earlier events caused Bill's action, they did so by affecting what he believed and wanted, which in turn caused him to act in a certain way.



Figure 18.5 Rose and Nichols' Incompatibilist Model. Arrows represent a direct, causal connection. The numbers above each edge are linear coefficients.

Throughpass questions are supposed to be the polar opposite of bypassing questions: they present a universe in which agents' mental states play a crucial role in the generation of their action. Thus, if bypassing questions are interpreted by participants as intended by Nahmias and Murray, we should observe an inverse correlation between participants' agreement to the throughpass and bypassing statements.

Björnsson's results raise two problems for the Bypassing Hypothesis. First, he found that bypassing scores did not fully mediate the effect of scenarios on FW/MR scores, which suggests that participants' tendency to confuse determinism with bypassing does not completely explain their incompatibilist judgments. Second, he found a weak but *positive* correlation between bypassing and throughpass statements, which suggests that participants might not interpret bypassing questions as intended by Nahmias and Murray.

Taken together, these findings cast doubt on the validity of the Bypassing Hypothesis. Should we then discard the Bypassing Hypothesis? I don't think so. Why? Here is the short answer: because, even if it faces some issues, it still constitutes a very good explanation for a range of phenomena, such as the difference between the *abstract* and *concrete* condition and the difference between psychological and neurological determinism. Rose and Nichols' Incompatibilist Model might well be a better fit of the data; it does not have this kind of explanatory power. In fact, it is not even clear whether there is a good psychological explanation that would correspond to the Incompatibilist Model. Why would judgments about free will and moral responsibility drive judgments about bypassing? Is it because participants first judge that free will and moral responsibility are impossible in a deterministic universe, then try to justify this judgment by appealing to bypassing-related considerations? But if participants were truly incompatibilists (and not compatibilists), they should not feel the need to point at such considerations to justify their judgment. Thus, it is not clear how we are supposed to make psychological sense of the Incompatibilist Model and, despite its problems, the Bypass Hypothesis is still the best available explanation for participants' judgments in a wider range of cases.

3.2 Answering the Throughpass Objection

Now, for those who would not be satisfied with this short answer, there is a longer, much more technical one. Let's first begin with Björnsson's most puzzling finding: that there is a *positive* correlation between bypassing and throughpass statements, that are supposed to measure opposed constructs. How are we to account for these results?

To explain them, I must introduce a distinction between the *deep* self (and deep motivational states) and the *superficial* self (and superficial motivational states). In his 1889 book on *Time and Free Will*, French philosopher Henri Bergson developed a conception of the human mind as composed of several layers (Bergson 1889/2002). The mental states that belong to the innermost layer constitute the "deep self": these are the values and commitments that

define who we *really* are. The outermost, shallow layers are constituted by mental states that are not truly “ours”: habits, automatism, social conditioning and pressure, etc. According to Bergson, to be free is to break the “crust” of shallow mental states to act on the basis of our “deep self.”

Similar ideas can be found in contemporary analytical philosophy (Watson 1975; Wolf 1993; Sripada 2016). More importantly, recent empirical studies have confirmed that most people naturally think this way, and distinguish between one’s deep self and one’s superficial self (Cova 2011; Sripada 2012; Newman, Bloom and Knobe 2014). Thus, people naturally make a distinction between two kinds of mental states: those who belong to the deep self, and reflect what we really care about, and those who belong to the superficial self, and can go against what we *really* believe and want.

With this distinction at hand, we have an easy explanation for Björnsson’s results. Take the following example: John really loves Sally. But Black, a master hypnotist hypnotizes John to induce in him the irrepressible desire to be mean and cruel to Sally. Now, John has two kinds of mental states: a desire to help and protect Sally (that reflect what he really cares about, and is thus part of his deep self) and a desire to harm her (that has been induced by Black, and is thus part of his superficial self). Now, imagine that we present participants with this case. We then ask them how much they agree with the corresponding throughpass statement:

When earlier events caused John’s action, they did so by affecting what he believed and wanted, which in turn caused him to act in a certain way.

Here, we can expect most participants to agree with this statement: after all, when Black causes John to be mean to Sally, he does so by affecting John’s mental states (i.e., desires). But, let’s now imagine that participants are presented with the following bypassing statement:

What John wants has no effect on what he ends up being caused to do.

Would participants agree with this statement? In one sense, they should disagree: if John is now mean to Sally, it is precisely because he wants to be. But, at the same time, there is a sense in which this statement is true: what John *really* wants (i.e., being kind to Sally) has actually no effect of what he ends up being caused to do. The statement is thus ambiguous: should “what John wants” be interpreted in a broad way, as encompassing his superficial desires and wants, or in a more restricted sense, as “what he *really* wants”?

Let’s now imagine that the majority of participants actually adopt the narrow interpretation, and think they are asked about what John *really* wants. In this case, we should expect agreement with the bypassing statement to be high. And, thus, we obtain a positive correlation between throughpass and bypassing statements. But this correlation does not constitute a contradiction: one can think that John’s actions are caused by his mental states, without accepting that they are caused by his *deep* mental states. The same is true for Björnsson’s results: if participants take determinism to imply that one is caused to act only by their superficial mental states, and thus that deep mental states play no role in the generation of one’s actions, we should also observe the positive correlation Björnsson observed.

Is this explanation of Björnsson’s results the right one? To find out, I conducted an online study in which participants were presented with one vignette introducing two characters: John, a gifted and ambitious neuroscientist who needs money, and Bill, the neighbor of John’s aunt. At the end of the vignette, John always made it so that Bill killed the aunt and that John inherited her money. The vignette existed in 6 different version, varying across two

dimensions. The first dimension was *Manipulation*: how John made it so that Bill ended up killing the aunt. In one case (*Body control*), John directly took control of Bill's body, not intervening on Bill's mental states. In another case (*Mind control*), John induced in Bill a strong and irresistible urge to kill John's aunt. Finally, in the third option (*Money*), John, knowing that Bill was in desperate need of money, simply offered Bill money in exchange of killing his aunt. The second dimension was *Reasons*: whether Bill had personal reasons *for* killing John's aunt (he hated her) or had personal reasons *against* killing her (they were friends). After reading the vignette, participants were asked to indicate their agreement (on a scale from -3 to 3) with a series of statements, including (i) statements about Bill's moral responsibility, statements about Bill's free will, (iii) bypassing statements, and (iv) a throughpass statement. The results of this study are presented in Table 18.2.

As expected, bypassing scores were very high for the *Body control* condition (in which the agents' mental states played no role) and low in the *Money* condition (in which the agent acts on his own). Bypassing scores in the *Mind control* condition (in which the agent acts on the basis of a mental state, but one that is externally induced) were significantly lower than in the *Body control* condition, but still higher than midpoint, and way higher than in the *Money* condition. This shows that most (but not all) participants interpret the bypassing statements as not asking about any kind of mental state, but about *deep* and/or *authentic* mental states.

Participants' ratings for the throughpass statements were puzzling. Even in the *Body control* condition, participants tended to agree with the throughpass statement. Moreover, compared to the *Body control* condition, throughpass ratings were not significantly higher in the *Money* condition. Even worse, their agreement with the throughpass statement was significantly higher in the *Mind control* condition than in the *Money* condition. As predicted, this suggests that the throughpass statement absolutely fails to measure what it was designed to measure (whether an agent's mental states play a crucial role in the formation of their action) and rather seems to measure to which extent the agent is acted by external forces. Thus, defenders of the Bypassing Hypothesis do not need to worry about the fact that throughpass and bypassing statements are not negatively correlated.

Before we move on, one interesting result should be noted: though above the midpoint, attributions of free will were quite low in the *Money* condition, compared to moral responsibility attributions. This is probably due to the fact that external circumstances (dire lack of money) put pressure on Bill to accept John's offer. Thus, some participants seem to use "free will" in a very demanding sense, in which one has to make decision independently from external pressures to count as "free." However, this sense is probably not the one relevant to theoretical theorizing, as most philosophical accounts of free will would certainly conclude

Table 18.2 Mean and SDs for participants' answers to the manipulation study in function of type of *Manipulation* and *Reasons* for responsibility, bypassing and throughpass scores.

	Body control		Mind control		Money	
Reasons	For	Against	For	Against	For	Against
Resp.	-2.15 (1.26)	-2.38 (1.28)	-1.08 (1.81)	-1.94 (1.32)	2.12 (1.09)	2.21 (0.99)
Free Will	-2.70 (0.94)	-2.77 (1.05)	-2.16 (1.52)	-2.74 (0.70)	0.58 (2.10)	0.55 (2.06)
Bypass	1.85 (1.38)	2.39 (1.07)	1.13 (1.46)	1.89 (1.29)	-1.44 (1.05)	-1.27 (1.15)
Throughpass	0.11 (2.13)	0.59 (2.39)	1.08 (1.65)	1.43 (1.88)	0.59 (1.59)	0.78 (1.53)
N	87	73	75	90	83	91

that Bill killed John's aunt of his own free will. Thus, this is one more argument in favor of the conclusion that moral responsibility ratings might do a better job at tracking the philosophically relevant concept of free will.

3.3 Answering the “Best-fit-of-the-data” Objection

Now, what about the fact that bypassing statements do not fully mediate the effect of scenarios on FW/MR judgments? My reply here is going to be technical but can be summarized as follows: this objection presupposes that, if participants' perceptions of bypassing explain the effect of determinism on their free will and moral responsibility judgments, then we should expect participants' agreement with bypassing statements to *fully* mediate the effect of determinism on their free will and moral responsibility judgments. However, this assumption is simply wrong.

More precisely, this assumption would be right if we were able to measure with complete accuracy participants' perceptions of bypassing, free will, and moral responsibility. But this is an unreasonable assumption, as items measures are susceptible to be interpreted in different ways by different participants. And, as soon as some measurement error is introduced, expectations of full mediations might no longer be reasonable.

To make this answer more salient, I decided to simulate some data. I distributed 200 virtual participants across two conditions: *Determinism* and *Indeterminism* (100 in each condition). Participants in the *Determinism* condition were randomly assigned a Bypassing score between 1 and 4, and those in the *Indeterminism* condition were randomly assigned a score between 4 and 7. Free Will scores were randomly computed from Bypassing scores by adding -1 , 0 or 1 to Bypassing scores. Thus, we simulate a causal chain such as: condition $\rightarrow$ bypassing $\rightarrow$ free will.

When we analyze these simulated data, we find as expected that the effect of condition on Free Will is fully mediated by Bypassing, but that the effect of condition on Bypassing is *not* fully mediated by Free Will. So, the results of mediation analysis accurately describe the underlying causal structure.

However, in this case, our measures *perfectly* represent the underlying phenomena. What happens when a little noise is added? To find out, I randomly added some noise to the data by randomly adding -1 , -0.5 , 0 , 0.5 , 1 to the Bypassing and Free Will scores. The new “inaccurate” scores were still strongly correlated with the original scores ($r = .92$ and $.93$ respectively), meaning that only a little measurement error was introduced. But this was already enough to make it so that Bypassing no longer *fully* (but only *partially*) mediated the effect of condition on Free Will (see Figure 18.6).

So much for the full mediation criterion, but what about choosing the model that best fits the data? To find out, I added more noise in Bypassing scores to simulate a situation in which the measurement error for Bypassing scores is greater than for Free Will scores (so that the correlation between original and actual Bypassing scores was 0.62). I then entered these data in the TETRAD program and asked the Greedy Search Algorithm to search for the model that best fitted the data. As can be seen in Figure 18.6, the model selected by the algorithm did not match the original causal structure. I also asked TETRAD to compare two causal models of the data: (i) condition $\rightarrow$ bypassing $\rightarrow$ free will, and (ii) (i) condition $\rightarrow$ free will $\rightarrow$ bypassing. It concluded that the second model was a way better fit than the first one. Thus, measurement error also compromises the use of such methods to determine which causal structure is the right one, particularly when the measurement error is higher for the mediator than for the variable we seek to predict.

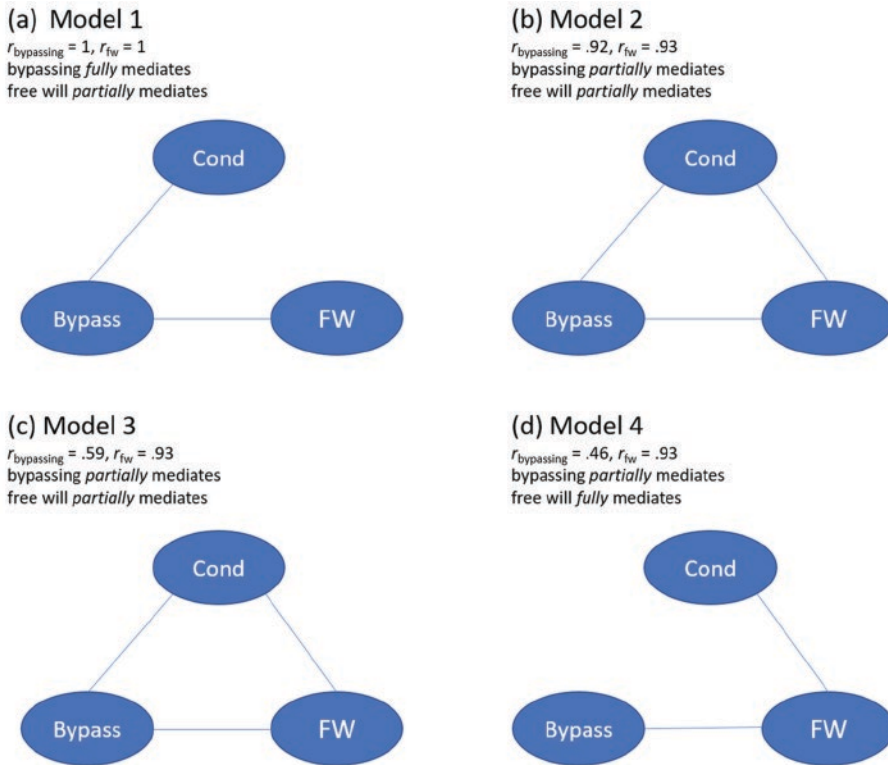


Figure 18.6 Causal models selected by the TETRAD Greedy Search Algorithm for simulated data, depending on the magnitude of measurement error. Measurement error is indicated by the correlation between actual data and original data (r).

Are we in such a situation? I think we can reasonably say “yes.” Participants’ judgments about free will and moral responsibility are precisely the outcome we seek to measure so, barring lies and inattention, their answer should accurately reflect the phenomenon we are interested in. But we saw that bypassing statements were subject to ambiguity: some participants took them as speaking of agents’ mental states in general (including externally induced mental states) while other interpreted them as asking about the agent’s *deep* and *authentic* mental states. This ambiguity naturally translates in measurement error. Thus, we are clearly in a situation where giving too much weight to the fact that there is not a *full* mediation or that a model best fit the data than another might lead us to make wrong theoretical choices.

Not that I am rejecting the use of mediation analysis as a *test* for a certain hypothesis: clearly, if a theoretical account predicts that a phenomenon M explains the connection between two other phenomena A and B , we should expect measures of M to mediate the link between measures of A and B . However, present measurement error, it might not be reasonable to demand that this mediation be a *full* mediation or this model to be the best fit of the data, as measurement error tends to underestimate the indirect effect and overestimate the direct effect (VanderWeele, Valeri, and Ogburn 2012). Such statistical considerations should not trump other arguments for the theoretical model, such as its ability to explain a whole range of phenomena.

4 An Error Theory for Compatibilist Intuitions

So far, I have argued in favor of the Bypassing Hypothesis, according to which seemingly incompatibilist answers are not genuinely incompatibilist but result from a confusion between determinism and bypassing. Thus, we could be tempted to conclude that, once that confusion is cleared, it turns out that most people have compatibilist intuitions. However, in the recent years, some have pushed the same kind of suspicions against compatibilist answers: participants' seemingly compatibilist answers might not be really compatibilists and could be explained away in a similar way we explained away incompatibilist answers. This approach is not incompatible with the Bypassing Hypothesis (it could that both seemingly compatibilist and incompatibilist answers can be explained away as mistakes), but it sheds doubt on the conclusion that laypeople are natural compatibilists.

4.1 *Early Error Theories*

Examples of such error theories for compatibilist answers are Feltz and Millan's (2015) claim that people will ascribe "free-will-no-matter-what," or Mandelbaum and Ripley's (2012) "Norm Broken, Agent Responsible" (NBAR) account, according to which participants will ascribe moral responsibility to agents as soon as they perceive that a norm is broken (which might explain why participants attribute moral responsibility in *concrete* cases but not in *abstract* ones). However, such accounts are not very plausible, since people do *not* automatically ascribe free will and moral responsibility, even when a norm is broken. For example, in Andow and Cova (2016), we asked participants to imagine a universe in which everyone's future is written in a magic book so that, anytime a person tries to act contrary to what is written in the book, the book will magically force this person to act against their will. Participants were presented with the story of John, whom the magical book forces to kill his wife against his will. In this case, most participants concluded that John did not act of his own free will, and was not morally responsible for killing his wife, showing that people do *not* ascribe free will "no matter what."

4.2 *Folks as Natural Incompatibilists: The Intrusive Metaphysics Account*

A more recent and subtle error theory is Nadelhoffer and colleagues' (2020) "intrusive metaphysics" account, according to which most participants fail to accept and integrate the deterministic assumptions of thought experiments such as the *Supercomputer* case but rather "import an indeterministic metaphysics" into their interpretation of these scenarios. To put it otherwise: the apparent compatibilist judgments of participants who seem to accept that agents in deterministic universes can be free and morally responsible for their actions could be explained away by their inability to truly accept the deterministic setting of the thought-experiment and their tendency to reinterpret it in indeterministic ways (see Figure 18.7).

To test this hypothesis, Nadelhoffer and his colleagues gave participants abstract and concrete versions of the *Supercomputer* and *Eternal Recurrence* cases developed by Nahmias and colleagues (2006). Half of participants received a version of these cases in which the agent was a robot rather than a human, but I will leave those aside to focus on participants' answers to cases involving human agent. In these cases, participants were asked the traditional questions about the agent's free will and moral responsibility, but also a series of questions aiming to probe to which extent they succeeded or failed to accept the deterministic setting of the scenario. For example, in the concrete version of the *Supercomputer* case, they were asked:

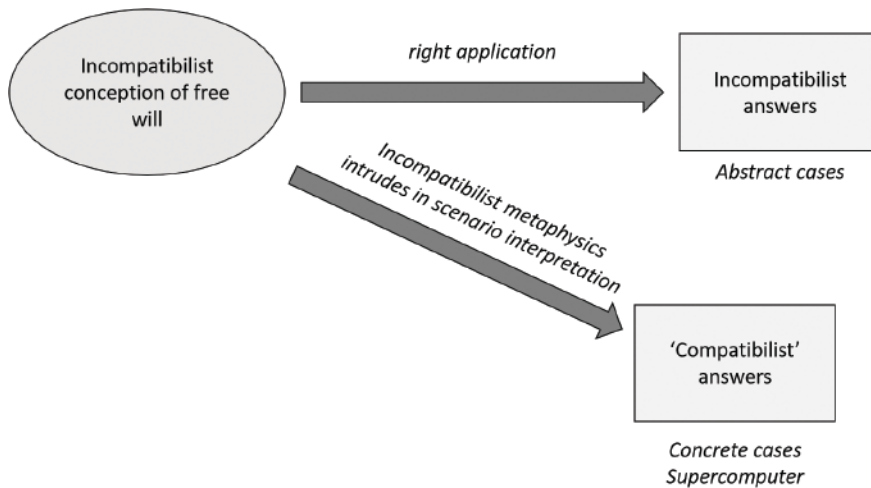


Figure 18.7 Nadelhoffer and colleagues' Intrusive Metaphysics Account.

Chance: What do you think the chances are that Jeremy will do something different than what the computer predicts he will do? (Slider scale ranging from 0 = very unlikely to 100 = very likely)

Here, I will focus on Chance as a measure of “metaphysical intrusion,” as Nadelhoffer and his colleagues themselves admit that “[Chance] is perhaps our best measure of intrusion.”

So what did they find? First, they found that a lot of participants indeed failed to integrate and accept the deterministic features of the vignettes. For *concrete* cases, only 47% of participants answered that the agent had *zero* chance to do something different. For *abstract* cases it was even less: 32%. Thus, it is clear that a lot of participants *do* import indeterministic assumptions in their understanding of these scenarios. This is a methodological issue future studies should keep in mind.

But what happens if we only keep participants who answered that the agent had *zero* chance to act in another way? Do they give overwhelmingly incompatibilist answers? As we can see in Table 18.3, there is no simple answer to this question. For the *abstract* cases, participants' free will and moral responsibility are very low – which is not surprising, as this is in line with previous results (see section 18.2.2). For the *concrete* cases, results are more mixed. Free will ratings (for the *free will* and *freely* questions) are split around the midpoint, with basically half of participants giving rather compatibilist answers and half of participants giving incompatibilist answers. However, when we look at questions about moral responsibility, blameworthiness/praiseworthiness, and desert, roughly two thirds of participants still give compatibilist answers (which is still pretty close to the results obtained by Nahmias and colleagues). If we take into account that the number of incompatibilist answers is likely to be inflated by the fact that Nadelhoffer and colleagues did not exclude participants who confused determinism with bypassing (since they had no bypassing items), then the results are not a big deviation from what has been observed by the previous literature, as long as we focus on moral responsibility (as I have argued we should, and as I intend to do in this chapter).

Thus, a true estimate of participants' intuitions about moral responsibility would require to exclude both seemingly “incompatibilist” answers (due to confusion with bypassing) *and*

Table 18.3 Percentage of answers *below* vs. *above* the midpoint for questions about free will and moral responsibility of human agents in Nadelhoffer et al. (2020), in function of whether participants answered that the agent had zero chance to act differently. For blame/praise and desert questions, answers were reverse-coded when the agents' action was good and I present the percentage of answers *below or equal* vs. *above* the midpoint. This re-analysis of Nadelhoffer et al.'s (2020) data was made possible by the fact that they shared their original data on osf.io/4z6r2/.

Chance	Concrete cases		Abstract cases	
	Chance = 0	Chance > 0	Chance = 0	Chance > 0
<i>of his own free will</i>	39% vs. 48%	14% vs. 78%	73% vs. 21%	13% vs. 78%
<i>freely</i>	40% vs. 47%	13% vs. 77%	71% vs. 24%	14% vs. 76%
<i>fully morally responsible for</i>	24% vs. 62%	11% vs. 79%	44% vs. 41%	08% vs. 82%
<i>is blameworthy/praiseworthy</i>	31% vs. 69%	10% vs. 90%	-	-
<i>deserves reward/punish</i>	31% vs. 69%	15% vs. 85%	-	-
N	212	238	70	148

seemingly “compatibilist” answers (due to intrusion of incompatibilist assumptions). Given the results described in Table 18.3, I do not think doing so will lead to results widely different from those already observed in the literature – for moral responsibility at least.

4.3 Additional Limitations of the Intrusive Metaphysics Account

Additionally, the Intrusive Metaphysics Account suffers from the same problem as the Incompatibilist Model: it has very low explanatory power. As we saw in Table 18.3, it cannot account for the difference between abstract and concrete cases, since this asymmetry subsists even when participants who make incompatibilist assumptions are excluded. Similarly, it is not clear that it can explain the difference between neurological and psychological determinism: why would participants import indeterministic assumptions in the second but not in the first case? Here, the account needs to be refined.

Moreover, the account claims that people import indeterministic assumptions into their comprehension of scenarios because “intuitive views about the indeterministic nature of human agency influence how people understand deterministic cases like *Supercomputer*.” However, this presupposes that people have an indeterministic view of human agency. But is it really the case? It is true that, when asked which of Nichols and Knobe’s Universe A and B is more like ours, most people choose Universe B (the indeterministic one). However, we have seen that participants’ understanding of these universes is plagued by a confusion between determinism and bypassing.

In a recent study, Giraud and Cova (in preparation; see also Monsieur Phi 2021) asked a total of 4430 volunteers to imagine that they had been victim of a strange physics experiment and that they were now condemned to live in a temporal loop, regularly being sent back to the same point in time. Then, participants were asked the following question:

Do you think the other persons around you in the temporal loop will always act the same way as long as you don’t intervene?

- Yes: they will always repeat the same behaviors
- No: certain behaviors will sometimes be different
- I can't say

In this case, 72.5% of participants answered “Yes,” that people in the loop will always repeat the same behaviors, which seems at odds with the idea that people have a strong indeterministic conception of human agency. Moreover, they were then asked to imagine that, in this loop, they witness two men (Bob and Charlie) repeatedly killing their respective neighbor. The sole difference was that Bob can easily be talked out of killing his neighbor, while Charlie can't. Among the participants who answered “Yes” to the first question (and accepted determinism), 79% answered that Bob acted freely and 91% answered that he was morally responsible for killing his neighbor. Moreover, 81% answered that Charlie acted freely and 92% answered that he was morally responsible. Thus, it seems that most people do not see human agency as indeterministic, and that seeing it as deterministic does not prevent them from attributing free will and moral responsibility. As such, one of the main assumptions of the Pervasive Metaphysics account still needs to be motivated.

5 A Final Question in Guise of Conclusion

In this chapter, I tried to give an overview of current debates on laypeople's conceptions of free will and moral responsibility and their relationship with determinism. I have tried to push forward the view that people tend to be natural compatibilists, as it seems to me the best approximation of truth. However, there are still too many controversies to present this conclusion as definitive. But I would like to end on one of the main reasons why I am attracted towards natural compatibilism: I understand why people would be natural compatibilists, but I don't understand why they would be natural incompatibilists. As mentioned earlier, practices and attitudes that presuppose the attribution of moral responsibility are pervasive. They also serve an important role in regulation social relationships. As such, it is reasonable that they evolved (biologically and/or culturally) to serve certain functions. However, I can't see how these functions would be best served by imposing incompatibilist demands on moral responsibility. Indeed, from a practical point of view, such demands seem pointless. Thus, I can't see *why* and *how* we would have come to develop an intuitive incompatibilist conception of moral responsibility in the first place. To me, this is one of the most important challenges defenders of natural incompatibilism need to face.

Supplementary Materials

Data for the study, simulation and re-analysis of Nadelhoffer et al.'s (2020) results presented in this paper can be found on the *Open Science Framework* at the following address: <https://osf.io/8aw5e>

Note

- 1 Of course, not every philosopher accepts this definition of free will (see for example van Inwagen 2008). However, for reasons that will get clearer as we progress, this is the one I will be using here. One objection to this kind of definition is that “moral responsibility” comes in different varieties (Rossi and Warfield 2016). In the context of this paper, I will use “moral responsibility” in the sense of *accountability* (Watson 1996).

Bibliography

- Andow, J. and Cova, F. (2016). Why compatibilist intuitions are not mistaken: a reply to Feltz and Millan. *Philosophical Psychology* 29 (4): 550–566.
- Bergson, H. (1889/2002). *Time and Free Will: An Essay on the Immediate Data of Consciousness*. Routledge.
- Berņiūnas, R., Beinorius, A., Dranseika, V., Silius, V., and Rimkevičius, P. (2021). The weirdness of belief in free will. *Consciousness and Cognition* 87: 103054.
- Björnsson, G. (2014). Incompatibilism and “bypassed” agency. In: *Surrounding Free Will* (ed. A. Mele), 95–122. New York: Oxford University Press.
- Björnsson, G. (2016). Outsourcing the deep self: deep self discordance does not explain away intuitions in manipulation arguments. *Philosophical Psychology* 29 (5): 637–653.
- Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., and Baumeister, R. F. (2014). Free to punish: a motivated account of free will belief. *Journal of Personality and Social Psychology*, 106(4): 501–513.
- Cova, F. (2011). *L'architecture de la cognition morale*. Ecole des Hautes Etudes en Sciences Sociales.
- Cova, F. (2014). Frankfurt-style cases user manual: why Frankfurt-style enabling cases do not necessitate tech support. *Ethical Theory and Moral Practice* 17 (3): 505–521.
- Cova, F. (2017). Frankfurt-style cases and the explanation condition for moral responsibility: a reply to Swenson. *Acta Analytica* 32 (4): 427–446.
- Cova, F. (forthcoming). “It was all a cruel angel’s thesis from the start”: folk intuitions about zygote cases do not support the zygote argument. In: *Advances in Experimental Philosophy of Free Will and Moral Responsibility* (ed. T. Nadelhoffer). London: Bloomsbury Press.
- Cova, F., & Kitano, Y. (2014). Experimental philosophy and the compatibility of free will and determinism: a survey. *Annals of the Japan Association for Philosophy of Science*, 22, 17–37.
- Cova, F., Bertoux, M., Bourgeois-Gironde, S., and Dubois, B. (2012). Judgments about moral responsibility and determinism in patients with behavioural variant of frontotemporal dementia: still compatibilists. *Consciousness and Cognition* 21 (2): 851–864.
- Doris, J.M., Knobe, J., and Woolfolk, R.L. (2007). Variantism about responsibility. *Philosophical Perspectives* 21: 183–214.
- Double, R. (1996). *Metaphilosophy and Free Will*. New York: Oxford University Press.
- Feltz, A. (2013). Pereboom and premises: asking the right questions in the experimental philosophy of free will. *Consciousness and Cognition* 22 (1): 53–63.
- Feltz, A. (2016). Folk intuitions. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), 468–476. New York: Routledge.
- Feltz, A. and Cova, F. (2014). Moral responsibility and free will: a meta-analysis. *Consciousness and Cognition* 30: 234–246.
- Feltz, A. and Millan, M. (2015). An error theory for compatibilist intuitions. *Philosophical Psychology* 28 (4): 529–555.
- Giraud, T. and Cova, F. (in preparation). Time loops and folk intuitions about free will. Unpublished manuscript.
- Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press.
- Knobe, J. and Nichols, S. (2010). Free will and the bounds of the self. In: *Oxford Handbook on Free Will*, 2e (R. Kane), 530–554. Oxford: Oxford University Press.
- Mandelbaum, E. and Ripley, D. (2012). Explaining the abstract/concrete paradoxes in moral psychology: the NBAR hypothesis. *Review of Philosophy and Psychology* 3 (3): 351–368.
- Mele, A.R. (2008). *Free Will and Luck*. New York: Oxford University Press.
- Miller, J.S. and Feltz, A. (2011). Frankfurt and the folk: an experimental investigation of Frankfurt-style cases. *Consciousness and Cognition* 20 (2): 401–414.
- Monsieur Phi. (2021). Les philosophes ne comprennent rien à la liberté. Retrieved the 25/ 27/2021 at <https://youtu.be/FuqIY-Xf5Is>
- Murray, D. and Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research* 88 (2): 434–467.

- Nadelhoffer, T., Rose, D., Buckwalter, W., and Nichols, S. (2020). Natural compatibilism, indeterminism, and intrusive metaphysics. *Cognitive Science* 44 (8): e12873.
- Nahmias, E. (2006). Folk fears about freedom and responsibility: determinism vs. reductionism. *Journal of Cognition and Culture* 6 (1–2): 215–237.
- Nahmias, E., Coates, D.J., and Kvaran, T. (2007). Free will, moral responsibility, and mechanism: experiments on folk intuitions. *Midwest Studies in Philosophy* 31 (1): 214–242.
- Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2005). Surveying freedom: folk intuitions about free will and moral responsibility. *Philosophical Psychology* 18 (5): 561–584.
- Nahmias, E., Morris, S.G., Nadelhoffer, T., and Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research* 73 (1): 28–53.
- Nahmias, E. and Murray, D. (2011). Experimental philosophy on free will: an error theory for incompatibilist intuitions. In: *New Waves in Philosophy of Action* (ed. J. Aguilar, A. Buckareff and K. Frankish). New York: Palgrave-Macmillan.
- Newman, G.E., Bloom, P., and Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin* 40 (2): 203–216.
- Nichols, S. and Knobe, J. (2007). Moral responsibility and determinism: the cognitive science of folk intuitions. *Noûs* 41 (4): 663–685.
- Rose, D. and Nichols, S. (2013). The lesson of bypassing. *Review of Philosophy and Psychology* 4 (4): 599–619.
- Rossi, B. and Warfield, T.A. (2016). The relationship between moral responsibility and freedom. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), 612–622, New York: Routledge.
- Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., and Sirker, S. (2010). Is belief in free will a cultural universal? *Mind & Language* 25 (3): 346–358.
- Sripada, C.S. (2012). What makes a manipulated agent unfree? *Philosophy and Phenomenological Research* 85 (3): 563–593.
- Sripada, C. (2016). Self-expression: a deep self theory of moral responsibility. *Philosophical Studies* 173 (5): 1203–1232.
- van Inwagen, P. (2008). How to think about the problem of free will. *The Journal of Ethics* 12 (3–4): 327–341.
- VanderWeele, T.J., Valeri, L., and Ogburn, E.L. (2012). The role of measurement error and misclassification in mediation analysis. *Epidemiology* 23 (4): 561.
- Watson, G. (1975). Free agency. *The Journal of Philosophy* 72: 205–220.
- Watson, G. (1996). Two faces of responsibility. *Philosophical Topics* 24 (2): 227–248.
- Wolf, S. (1993). *Freedom within Reason*. New York: Oxford University Press.

Libertarianism¹

MARK BALAGUER

DEPARTMENT OF PHILOSOPHY, CALIFORNIA STATE UNIVERSITY, LOS ANGELES

This chapter is about the libertarian view of free will.² I'll start by saying a few words about how libertarianism ought to be defined. Then I'll formulate some classic arguments against libertarianism, and finally, I'll develop a version of event-causal libertarianism and argue (very briefly) that if we adopt this version of the view then we can respond to the arguments against libertarianism.

1 Defining Libertarianism

Libertarianism has traditionally been defined as the conjunction of two theses, namely, (a) the thesis that human beings have free will, and (b) *incompatibilism* – i.e., the thesis that free will is incompatible with determinism. But this definition doesn't do a very good job of capturing everything that libertarians are committed to (and I think it also fails to get at the *bottom-level* commitments of libertarians). The first step toward getting a better definition is to articulate the kind of freedom that libertarians believe in. We can do this as follows:

A person is *libertarian-free* (or for short, *L-free*) if and only if she makes at least some decisions that are such that (a) they are both undetermined and appropriately non-random, and (b) the indeterminacy is relevant to the appropriate non-randomness in the sense that it *generates* the non-randomness, or *procures* it, or *enhances* it, or *increases* it, or something along these lines.

More needs to be said about what appropriate non-randomness consists in. Different libertarians will say different things about this, but all should agree that it consists in some sort of *agent-involvedness*. For example, one might say that it consists in the agent *controlling* which options is chosen, or *authoring* the choice, or being the *source* of the choice, or making a *rational* choice, or some combination of these things. Also, many libertarians would follow Kane (1985, 1996) in requiring *plural* control (or authorship or whatever) – i.e., in requiring it to be the case that even if the agent had chosen differently, she still would have controlled it or authored it, or whatever.

A Companion to Free Will, First Edition. Edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White.
© 2023 John Wiley & Sons, Inc. Published 2023 by John Wiley & Sons, Inc.

Also, more needs to be said about clause (b) of the definition of L-freedom. I intend this clause to be read very loosely; the idea is that the indeterminacy has to be relevant to the non-randomness (i.e., to the agent-involvedness) in *some way or other*. Different libertarians can fill this in in different ways. Some might say that indeterminacy is a *necessary condition* for non-randomness; others might deny this and maintain that, in some cases, the fact that a decision is undetermined somehow *causes* that decision to be non-random. I'll say more below about what I think libertarians should say about this, but for now, I want to leave it open.

Given the above definition of L-freedom, we can define libertarianism as the conjunction of the following two theses:

HB-libertarianism: Human beings are L-free.

C-libertarianism: Free will is L-freedom.^{3,4}

According to this new definition, libertarianism is a stronger view than it is on the traditional definition. To appreciate this, notice first that this new view (i.e., the conjunction of HB-libertarianism and C-libertarianism) clearly entails both conjuncts of the traditional view – i.e., that humans have free will and that incompatibilism is true. But the new view commits to much more than the traditional view does. According to the traditional definition, libertarians are committed to the thesis that human beings have an indeterministic kind of free will. But L-freedom isn't just indeterministic; there are further requirements that need to be met for an indeterministic kind of free will to count as a kind of L-freedom; in particular, the indeterminacy needs to generate (or procure, or whatever) an appropriate sort of non-randomness, or agent-involvedness.

Given that my definition takes libertarianism to be a stronger view than the traditional definition does, I should say a few words to motivate the idea that libertarians really are committed to what I'm saying they're committed to. The first point to note here is that libertarians are clearly committed to the thesis that human beings have a kind of free will that involves appropriate non-randomness, or agent-involvedness. If all of my decisions are controlled by Martians via remote control, then I don't have free will in the libertarian sense of the term (or any other sense, for that matter), and this is true even if there are indeterminacies in the causal pathways from the Martians' manipulations to my decisions. So indeterminacy is clearly not enough for libertarian-freedom; it also needs to be the case that our decisions are appropriately non-random – i.e., they need to be made *by us*. Second, it's also clear that libertarians are committed to the idea that the non-randomness (or agent-involvedness) has to be generated (or procured, or enhanced, or whatever) by the indeterminacy. Without this, the following kind of freedom would seem to count as a kind of libertarian-freedom:

A person is *virtually-Hume-free* iff (i) there are insignificant indeterminacies in her decision-making processes—i.e., indeterminacies that almost never have any effect on which options are chosen; and (ii) her decisions are “for-all-practical-purposes determined” by her desires—i.e., they're probabilistically caused by her desires with a high degree of certainty (in particular, the objective moment-of-choice probabilities of her choices are always at least 0.9999999).

If we're virtually Hume-free, then our decisions are undetermined, and they also seem to be non-random in an agent-involving way (because they're “virtually determined” by our desires, and so they *come from us* in an obvious way⁵). But virtual-Hume-freedom is clearly not a *libertarian* sort of freedom, and the reason is that the indeterminacy here is entirely irrelevant to the non-randomness. Indeed, if virtual-Hume-free decisions are non-random – i.e., if they're made *by us* – then this is *in spite of* the fact that they're undetermined. Libertarians

are clearly committed to something stronger than this; they're committed to the idea that the indeterminacy is *relevant* in some way to the non-randomness – i.e., to the idea that the indeterminacy generates the non-randomness, or procures it, or enhances it, or some such thing.

So I think that my definition of libertarianism is better than the traditional definition because it's more informative – i.e., it does a better job of capturing what libertarians are actually committed to. Moreover, my definition also does a better job of capturing the *bottom-level* commitments of libertarians. To appreciate this, notice that while the two conjuncts of the traditional definition are not independent of one another, the two conjuncts of my definition *are* independent of one another. One of the conjuncts (HB-libertarianism) is a contingent claim about the nature of the actual world, in particular, human decision-making processes; and the other conjunct (C-libertarianism) is a *conceptual* claim – it's an answer to the conceptual-analysis question 'What is free will?' But the answer to the conceptual-analysis question is clearly not relevant to the question of whether HB-libertarianism is true because the term 'free will' doesn't appear at all in the definition of HB-libertarianism. Moreover, *prima facie*, it seems that the answer to the contingent, actual-world question of whether HB-libertarianism is true isn't relevant to the conceptual-analysis question either. Now, you might try to resist this last claim by arguing that 'free will' should be defined as the freedom-like ability that humans actually have – whatever that turns out to be. But I think it can be argued that (a) this is in fact *not* the right way to define 'free will,' and (b) even if it is, the question of whether HB-libertarianism is true is still not relevant to the conceptual-analysis question 'What is free will?'⁶

2 Zeroing in on HB-Libertarianism

The question of whether libertarianism is true reduces to the following two questions:

The HB-question: Is HB-libertarianism true?

The C-question: Is C-libertarianism true?

These are both difficult, multi-faceted questions, and I don't have the space to do justice to both of them here. Thus, in what follows, I'm going to focus on the HB-question and ignore the C-question. I do this because I find the HB-question much more interesting. The reason I'm less interested in the C-question is that I think the answer to this question is (a) entirely settled by empirical facts about the usage and intentions of ordinary folk concerning words like 'free,' 'can,' 'could have,' etc., and (b) not relevant in any non-trivial way to metaphysical questions about the nature of human decision-making processes. Both of these claims are extremely controversial; I can't give the complete argument for them here, but I'd like to say a few words to clarify my view.⁷

The first point to notice here is that there are multiple kinds, or concepts, of freedom. For instance, there's L-freedom (defined above); and Hume-freedom (i.e., roughly, the ability to do what you want, or to act and choose in accordance with your desires); and Frankfurt-freedom (i.e., roughly, the ability to control, with second-order attitudes, which of your first-order desires will affect your behavior); and so on. (There are also multiple kinds, or concepts, of moral responsibility; e.g., there's L-responsibility (which requires L-freedom), and Hume-responsibility (which requires Hume-freedom), and Frankfurt-responsibility (which requires Frankfurt-freedom), and so on.) Now, in themselves, these are all perfectly good concepts. If you have Platonist leanings, you can think of them as existing "side-by-side" in Platonic heaven; if you don't have such leanings, you can tell whatever side-by-side story you

like. Either way, the point is that these are all perfectly good concepts that a community of rational, moral creatures could usefully employ.

(It's important to note that I'm not yet saying anything controversial. By merely claiming that these concepts exist, and that we can introduce terms of art that express them, all I'm really saying is that people can introduce new terms and define them however they like. For instance, if we wanted to, we could introduce the term 'one-footed freedom' to denote the ability to make a decision while standing on one foot. One-footed freedom isn't a very interesting kind of freedom, and it's clearly not the kind of freedom that's picked out by the English expression 'free will,' but there's nothing *wrong* with it, as a concept. So when I say that L-freedom and Hume-freedom and so on are "perfectly good concepts that a community of rational creatures could employ," I don't mean to suggest that any of these concepts captures the correct definition of the English expression 'free will.' Moreover, I also don't mean to suggest that these concepts are all instantiated, or even that they *could* be instantiated. As we'll see in Section 4, some people think that L-freedom is *incoherent* and that it's *impossible* for a person to be L-free. But even if this were true, it wouldn't follow that there was no such thing as the concept of L-freedom. Indeed, the claim that L-freedom is incoherent straightforwardly *depends on* the claim that there is a concept of L-freedom; for if there were no such concept, then there wouldn't be anything there to be incoherent, and the sentence 'L-freedom is incoherent' wouldn't be true – indeed, it would be gibberish.⁸)

In any event, given that we've got all of these concepts to work with, when philosophers ask the conceptual-analysis question 'What is free will?,' we can understand them as asking this: Which of the various concepts of freedom – or if you'd rather, which of the various *freedom-like* concepts, e.g., L-freedom, Hume-freedom, etc. – is identical to the concept of free will? In other words, they're asking which of these concepts provides the correct analysis of the concept of free will, or the correct definition of 'free will.' (Likewise, when philosophers ask what moral responsibility is, they're asking which of the various concepts of responsibility – or which of the various *responsibility-like* concepts – is identical to the concept of moral responsibility; i.e., they're asking which of these concepts provides the correct analysis of the concept of moral responsibility.) But what sorts of questions are these? What kinds of facts could determine the answers to these questions? After all, as I just pointed out, all of the concepts of freedom and responsibility are, in themselves, perfectly good. And it's not as if some of them are glowing with a special hue in Platonic heaven. So what could make two of these concepts stand out as *the* concepts of free will and moral responsibility?

One view here is that the relevant facts – the facts that determine which concepts count as the concepts of free will and moral responsibility – are facts about *us*. In particular, on this view, the concept of free will is just the concept that *we have in mind* when we think and talk about free will, and the concept of moral responsibility is the concept that we have in mind when we think and talk about moral responsibility. I think there are good reasons to think that (a) this view of conceptual analysis is true, and (b) it implies that the C-question is not relevant in any non-trivial way to questions about the nature of human decision-making processes.⁹ Now, of course, both of these claims require argument, and unfortunately, I can't say any more about this topic here, but this at least clarifies why I think the C-question is less interesting than the HB-question.

3 Arguments for HB-Libertarianism?

I don't think there are any good arguments for HB-libertarianism. Some might think that the fact that we have an "intuition" that we're free gives us good reason to believe that we're L-free, but

this strikes me as highly immodest. HB-libertarianism requires the truth of indeterminism, and indeterminism involves a substantive, controversial thesis of subatomic physics. The idea that our intuitions could track the truth about this seems to me about as plausible as the idea that our intuitions could track the truth about the number of planets that are orbiting Alpha Centauri.

One might also try to argue for HB-libertarianism from the following two claims: (i) we're morally responsible for our actions, and (ii) moral responsibility requires L-freedom.¹⁰ But the idea that moral considerations can give us good reason to believe a controversial thesis of sub-atomic physics is no more plausible than the idea that our intuitions can give us good reason to believe such a thesis. To say a bit more about this, I think the problem with the argument from (i) and (ii) is that while both of these premises can seem plausible when considered in isolation, it's not plausible to suppose that we currently have good reason to believe that they're *both* true. In particular, if we had good reason to believe thesis (ii), that would immediately give us good reason to think that we had no reason to believe thesis (i), for in this scenario, the truth of (i) would depend on a controversial thesis of subatomic physics.

4 Arguments Against HB-Libertarianism

Prima facie, the arguments *against* HB-libertarianism seem much stronger than the arguments in favor of that view. Here are three arguments against HB-libertarianism that, I think, a lot of people find convincing:

- (1) *The argument from determinism*: HB-libertarianism is false because it requires a certain sort of indeterminism, and there are good empirical reasons to think that this sort of indeterminism isn't true.
- (2) *The luck argument (aka, the Mind argument, or the argument from randomness, or the Hobbes-Hume-Hobart argument)*: Even if the relevant sort of indeterminism is true, it doesn't matter because there's a clear sense in which undetermined events are *random*, or *lucky*. In other words, they occur by *chance* – i.e., they *just happen*. Thus, if we introduce an undetermined event into a decision-making process, that would seem to either (a) increase the level of randomness (or luckiness) in that process or (b) leave the level of randomness (or luckiness) alone (if the indeterminacy ends up not mattering). So it's hard to see how the introduction of an undetermined event into a decision-making process could increase *non-randomness* (or non-luckiness, or agent-involvedness, or whatever). Thus, since this is precisely what HB-libertarianism requires, it seems that HB-libertarianism is false and, indeed, arguably incoherent.
- (3) *The argument from naturalism*: HB-libertarianism is incompatible with a naturalistic, scientifically reputable view of the world. In particular, it seems to be incompatible with a materialistic view. Therefore, since materialism is true, HB-libertarianism is false. (This argument is almost never discussed in the literature, but it's, so to speak, "in the air" – and, I think, in lots of people's heads.)

Among these three arguments, the luck argument has received the most attention, especially recently. Numerous versions of this argument have appeared in the literature, spread out over hundreds of years, under many different names.¹¹ There are also many responses to this argument.¹² In fact, I think it's fair to say that the various different versions of libertarianism have been developed in connection with attempts to respond to the luck argument. The most

important distinction here is between *agent-causal* views, *event-causal* views, and *non-causal* views. Agent-causal libertarians maintain that when humans make L-free decisions, they agent-cause these decisions to occur, where agent-causation is a kind of causation that obtains not between two events (as is the case with event-causation) but between an agent and an event (and according to agent-causal views, agent-causation is not reducible to event-causation).¹³ Event-causal libertarians, on the other hand, maintain that all causation reduces to ordinary event-causation and that our L-free decisions are probabilistically caused by agent-involving events.¹⁴ And non-causal libertarians maintain that our L-free decisions are uncaused.¹⁵ These three views have been developed, by and large, as ways of getting around the luck objection. I don't have the space to discuss all of them here, so I will focus on event-causal libertarianism. Indeed, I'll focus on one specific version of event-causalism, namely, the one that I think provides the right response to the luck argument.¹⁶

I'll spend most of the rest of this chapter responding to the luck argument, but at the end, I'll say a few words about the arguments from determinism and naturalism. Indeed, my response to the luck argument will bring with it a response to the argument from naturalism; for the version of libertarianism that I'll be developing here is entirely materialistic and scientifically reputable. In particular, I'll be assuming that human decisions are *neural* events.

(In addition to being materialistic and event-causal, the version of libertarianism I'll be developing here is *centered* – i.e., it places the important indeterminacy at the moment of choice. This is in contrast to *deliberative* libertarianism, which places the important indeterminacy *prior* to the moment of choice. Deliberativism has been developed by Dennett (1978) and Mele (1995), but to the best of my knowledge, no one has ever *endorsed* this view.¹⁷ This, I think, is for good reason; I think there are serious problems with deliberativism.)

5 Responding to the Luck Argument

5.1 Torn Decisions

In order to respond to the luck argument, it's important for libertarians to focus on decisions of the right kind – namely, what I have elsewhere called *torn decisions*. The notion of a torn decision can be defined as follows:

A *torn decision* is a decision in which the person in question has reasons for multiple options, feels torn as to which option is best (and has no conscious belief as to which option is best), and decides without resolving the conflict, i.e., decides while feeling torn.

I think we make decisions like this several times a day about things like whether to have eggs or cereal for breakfast, or whether to work out before leaving for work, or whatever. But we can also make torn decisions in connection with big life-changing decisions; e.g., you might have a good job offer in a city you don't like, and you might have a deadline that forces you to decide while feeling utterly torn.

Torn decisions should be distinguished from three other kinds of decisions. First, they should be distinguished from *leaning decisions*; these are decisions in which the agent chooses while leaning toward one of her live options, whereas in a torn decision, the agent feels *completely* torn. Second, torn decisions should be distinguished from *Buridan's-ass decisions*; these are similar to torn decisions except that the various tied-for-best options are more or less indistinguishable, and because of this, the agent *doesn't feel torn*. (For instance,

if you want a can of tomato soup, and there are ten cans of the same kind on the shelf, you won't feel torn – you'll just grab one and be on your way.¹⁸) Third, torn decisions should be distinguished from what Kane (1996) calls *self-forming actions*, or SFAs. The most important difference here is that, unlike SFAs, torn decisions are not defined as being undetermined. They're defined in terms of their phenomenology. Thus, we know from experience that we do make torn decisions, and it's an empirical question whether any of these decisions are undetermined.

My reason for thinking that libertarians should focus on torn decisions is not that I think that decisions of these other kinds can't be free, or L-free. My reason is just that if we focus on torn decisions, then, as we'll see in what follows, we can get a very clear response to the luck argument.

5.2 TDW-indeterminism

The next thing we need to do, in order to get a response to the luck argument, is to get clear on the relevant kind of indeterminacy – i.e., the kind of indeterminacy that needs to be present in torn decisions in order for decisions of this kind to be L-free. This sort of indeterminacy can be defined as follows:

A torn decision is *wholly undetermined* at the moment of choice—or, for short, *TDW-undetermined*—iff the actual objective moment-of-choice probabilities of the various reasons-based tied-for-best options being chosen match the reasons-based probabilities (or the phenomenological probabilities), so that these moment-of-choice probabilities are all roughly even, given the complete state of the world and all the laws of nature, and the choice occurs without any further causal input, i.e., without anything else being significantly causally relevant to which option is chosen.

It's important to note that this sort of indeterminacy is compatible with various features of the decision being fully determined. Suppose, e.g., that I'm about to make a torn decision between options A and B. It could be determined that (i) I'm going to make a torn decision (i.e., I'm not going to refrain from choosing); and it could be determined that (ii) I'm going to choose between A and B (i.e., I'm not going to choose some third option that I don't like as much); and it could be determined that (iii) the moment-of-choice probabilities of A and B being chosen are both 0.5 (i.e., they're not 0.7 and 0.3, or 1 and 0, or anything else). All of this is perfectly consistent with the decision being TDW-undetermined. All that needs to be undetermined, in order for the choice to be TDW-undetermined, is *which tied-for-best option is chosen*.

It's also important to note that TDW-indeterminacy lies at one end of a spectrum of possible cases and that there are *degrees* of the kind of indeterminacy I'm talking about here. To see what I've got in mind by this, suppose that Ralph makes a torn decision to order chocolate pie instead of apple pie. Since this is a torn decision, we know that given all of Ralph's conscious reasons and thought, he feels completely neutral between his two tied-for-best options. But it could be that factors external to Ralph's conscious reasons and thought (e.g., unconscious compulsions, or wholly non-mental brain events that precede the decision in his head) causally influence the choice and wholly or partially determine which option is chosen. Indeed, there's a continuum of possibilities here. At one end of the spectrum, which option is chosen is TDW-undetermined, so the moment-of-choice probabilities of the two tied-for-best options being chosen are 0.5 and 0.5, and nothing else significantly causally influences which option is chosen. At the other end of the spectrum, the choice is fully

determined – i.e., factors external to Ralph’s conscious reason and thought come in and, unbeknownst to Ralph, cause him to choose chocolate. And in between, there are possible cases where the moment-of-choice probabilities are neither 0.5 and 0.5 nor 1 and 0 – i.e., where they’re 0.8 and 0.2, or 0.7 and 0.3, or whatever; in these cases, external factors causally influence the choice without fully determining it. In cases like this, we can say that which option is chosen is *partially determined and partially undetermined*.¹⁹

5.3 The Central Libertarian Thesis

Let *TDW-indeterminism* be the view that some of our torn decisions are TDW-undetermined. Given this, I think that libertarians can respond to the luck argument by arguing for the following thesis:

Central Libertarian Thesis (CLT): If our torn decisions are undetermined in the right way—i.e., if they’re TDW-undetermined—then they’re appropriately non-random and L-free. Or more succinctly: *If TDW-indeterminism is true, then HB-libertarianism is true.*

Notice that CLT doesn’t just say that our torn decisions *could be* both undetermined and non-random; it says that if they’re undetermined in the right way, then they *are* appropriately non-random and L-free, so that the question of whether our torn decisions are L-free just reduces to the question of whether they’re undetermined in the right way. If this is right, then it doesn’t just refute the luck argument – it turns that argument completely on its head. For whereas the luck argument suggests that indeterminacy leads to randomness, CLT says the exact opposite – that the right kind of indeterminacy leads to *non-randomness*. This is much stronger than what other libertarians (e.g., Ginet (1990) and Kane (1996)) have tried to establish in responding to the luck argument; they’ve tried to show that indeterminacy doesn’t necessarily lead to non-randomness – i.e., that there are ways in which a decision could be undetermined and yet still be appropriately non-random. But, again, CLT says that if our torn decisions are TDW-undetermined, then they *in fact are* appropriately non-random and L-free. If this is right, then it gives us the surprising result that we can establish libertarianism – or at any rate, HB-libertarianism – by arguing for an *empirical* thesis, namely, TDW-indeterminism.

(CLT is only about *full-blown* L-freedom. I think it can also be argued that if an ordinary torn decision is partially undetermined in the manner described in Section 5.2, then it’s *partially* L-free. But I won’t worry about this here; I’ll just focus on arguing that if our torn decisions are *wholly* undetermined – i.e., TDW-undetermined – then they’re *fully* L-free.)

5.4 An Argument for CLT

I argued at length for CLT in my (2010). I will here just provide a brief formulation of one of the arguments. To get started, let’s return to Ralph’s decision to order chocolate pie instead of apple pie, and let’s suppose that he was completely torn when he made this decision. Given this, we get the following result:

(A) Ralph’s choice was conscious, intentional, and purposeful, with an actish phenomenology—in short, it was a Ralph-consciously-choosing event, or a Ralph-consciously-doing event.

If we now assume that Ralph’s decision was TDW-undetermined, then we also get the following two results:

- (B) Ralph's choice flowed out of his conscious reasons and thought in a nondeterministically event-causal way.
- (C) Nothing external to Ralph's conscious reasons and thought had any significant causal influence (at the moment of choice—i.e., after he moved into a torn state and was about to make his decision) over how he chose, so that the conscious choice *itself* was the event that settled which option was chosen.

My argument for CLT is based on the observation that (A)-(C) seem to imply that Ralph authored and controlled his decision. This is because (A)-(C) seem to give us the two-fold result that (i) *Ralph did it*, and (ii) *nothing made him do it*; and, intuitively, it seems that (i) and (ii) are sufficient for authorship and control. The Ralph-did-it-ness comes from (A)-(B) – the choice flowed out of Ralph's reasons, and the event that actually settled which option was chosen was a conscious Ralph-choosing event. And the fact that nothing made him do it comes from (C).

(It's important to note that I'm not engaged here in conceptual analysis. I'm not putting forward a theory of what authorship and control *are*. It might seem that when I say that (i) and (ii) are sufficient for authorship and control, I'm making a claim of conceptual analysis; but as will become clear below, all I'm committed to saying here is that in connection with ordinary torn decisions like Ralph's, if they're TDW-undetermined, then they're authored and controlled by the agent *in certain interesting and important ways*. I don't need to claim that the kinds of authorship and control that are present here are the only kinds of authorship and control that one might care about. And I don't need to say anything about the nature of authorship and control. More on this below.)

In any event, to get the result that Ralph's decision was appropriately non-random and L-free, it's not enough to argue that he authored and controlled his decision; we also need to argue that (i) his decision satisfied all of the other conditions for appropriate non-randomness, aside from authorship and control (e.g., rationality, the Kanean plurality conditions, and so on), and (ii) the fact that Ralph's decision was TDW-undetermined *procured* the result that it was appropriately non-random.

I don't have the space to say very much in support of these two claims here, but given the above argument about authorship and control, they're both relatively uncontroversial. For instance, regarding (i), it should be clear that Ralph's choice was extremely rational; if you have to choose between two options, and if you're completely unsure which option is best, and if not choosing (or continuing to deliberate) is worse than choosing arbitrarily, then the most rational thing to do is to arbitrarily pick one of your tied-for-best options. Moreover, given that Ralph was *torn*, it also seems clear that his decision satisfied the Kanean plurality conditions; for if he'd chosen to order apple pie instead of chocolate pie, we would have been able to run the same argument about authorship, control, rationality, and so on.

As for point (ii), this should already be clear from the argument about authorship and control. For the fact that Ralph's decision was TDW-undetermined played a crucial role in the argument for the claim that he authored and controlled the decision—this is what gave us the result that nothing made Ralph choose chocolate pie over apple pie. Moreover, as we'll shortly see, the fact that Ralph's decision was TDW-undetermined also gives us the result that the event that actually settled which option was chosen was the conscious decision *itself*. (The idea here isn't that TDW-indeterminacy actively *generates* authorship and control; the idea is that it *blocks a destroyer* of authorship and control. The destroyer would be a moment-of-choice causal influence from something external to Ralph's conscious reasons and thought.

TDW-indeterminacy *guarantees the absence* of such a destroyer, and this is the sense in which it procures authorship and control.)

If all of this is right, then we have an argument for CLT – although, I admit, it’s really just a sketch of an argument. The key to the argument, in my opinion, is that it gives us the following result:

Choice-Settles-It: The event that settled which option was chosen was the conscious choosing event.

You might object to this in two different ways. First, you might claim that if we’re assuming a materialistic view of the mind-brain, then the event that settled which option was chosen was a *neural* event. And second, you might claim that which option was chosen was settled by some *quantum* events. I agree with both of these claims, but I don’t think that either of them undermines Choice-Settles-It. For (a) I think that on any reasonable version of mind-brain materialism, the conscious choosing event *is* a neural event;²⁰ and (b) I think that the neural event in question is *composed* of quantum events, so that if Ralph’s decision was TDW-undetermined, then which option was chosen was settled by the quantum events that *made it up*. (If the decision was settled by *prior-to-choice* quantum events, then it wouldn’t be correct to say that it was settled by the conscious choice *itself*; but TDW-indeterminism rules out the possibility of this prior-to-choice determination; if the decision was TDW-undetermined, then it was settled by quantum events that were *parts of* the decision, not by quantum events that preceded the decision.)

The argument I’ve been rehearsing here differs from other libertarian arguments in the literature in numerous ways, but perhaps the most important difference has to do with how modest its assumptions are. Other libertarians have tended to respond to the luck argument by constructing complicated libertarian models and arguing that for all we know, these elaborate models could be true. I’ve taken the exact opposite approach here. I’ve argued that by assuming *only* that our torn decisions are TDW-undetermined, we get the result that the undetermined physical events that settle the outcomes of our torn decisions *are the decisions themselves* – i.e., the conscious mental events with a me-choosing-now phenomenology. And this, I think, gives us a clear sense in which the outcomes of these decisions are settled by us.

5.5 An Objection

The argument for CLT that I just articulated was obviously very quick, and there are a number of different worries that one might have about it. In this section, I’ll say a few words about one of these objections, namely, the following:

The Agent-Causal Objection: The notion of control that you’re working with is too weak. Something more substantive is needed for control. In particular, it seems that something like agent causation is needed. In other words, when someone makes a torn decision, in order for it to be the case that the agent in question controls which option is chosen, it needs to be that she *causes* the given option to be chosen. But if Ralph’s decision was TDW-undetermined, then he didn’t cause chocolate pie to be chosen; in fact, in this scenario, *nothing* caused chocolate pie to be chosen. Or at any rate, nothing deterministically caused it to be chosen. We can say that Ralph’s reasons probabilistically caused chocolate pie to be chosen, but there seems to be a clear sense in which nothing caused the choice to be a *chocolate-rather-than-apple* choice. And it seems in particular that *Ralph* didn’t cause this. And so we can’t correctly say that he *controlled* which option was chosen. (Pereboom raises a worry like this about my view in his (2014).)

One way to respond to this objection would be to argue that agent causation is in fact *not* required for authorship and control. I say a few words to motivate this stance in my (2014b), but the point I want to make here is that in the present context, it doesn't matter whether agent causation is required for authorship and control. To bring this out, let me distinguish two different kinds of control – *causal-control* and *noncausal-control* – where the former requires agent causation (or something like it) and the latter doesn't. I won't try to give precise definitions of these two notions; all I'll say (and all that will matter here) is that when I use the term 'noncausal-control,' I'm talking about a kind of control that applies to ordinary torn decisions if they're TDW-undetermined; i.e., it applies to torn decisions like Ralph's, where the agent makes a conscious decision with an actish phenomenology, and which option is chosen isn't significantly causally influenced (at the moment of choice) by anything external to the agent's conscious reasons and thought, so that the conscious choice itself *is* the event that settles which option is chosen. Beyond this (and beyond the fact that causal-control requires agent causation and noncausal-control doesn't), it won't matter how exactly these two kinds of control are defined. But for the sake of argument, let's pretend that we've got two precisely defined kinds of control here. Given this, one question we might ask is the following:

The what-is-control question: What is control? In other words, which of the various kinds of control that we find in the literature provide correct analyses of the concept of control? Does causal-control? Does noncausal-control? Do both? Do neither?

But why should libertarians care about this question? They don't need to claim that if our torn decisions are TDW-undetermined, then they're authored and controlled by us and L-free in the only senses of these terms that anyone might care about, or in the senses of these terms that philosophers have traditionally cared about. All they need is this:

(W) If our torn decisions are TDW-undetermined, then they're authored and controlled by us and appropriately non-random and L-free in interesting and important ways that are worth wanting and worth arguing for and that libertarians can hold up and say, "This gives us a noteworthy kind of libertarian free will."

Now, don't take me to be saying more than I am here. I don't think that libertarians can just define authorship and control and L-freedom in ridiculously weak ways and then claim victory. Or to put the point differently, while I don't need to argue that the kind of L-freedom I've articulated – the kind that we get if our torn decisions are TDW-undetermined (i.e., the kind that involves noncausal-control) – is the one and only kind of L-freedom that anyone might care about, I *do* need it to be the case that this kind of L-freedom is interesting, worth wanting, worth arguing for, and so on. In other words, I need (W).

But I think the above argument for CLT does motivate (W). Let's return to Ralph's decision. If it was TDW-undetermined, then (a) the choice was conscious, intentional, and purposeful, with an actish phenomenology; and (b) it flowed out of Ralph's conscious reasons and thought in a nondeterministically event-causal way; and (c) nothing external to Ralph's conscious reasons and thought had any significant causal influence (after he moved into a torn state and was about to make his decision) over how he chose. This might not give us every kind of L-freedom you might have wanted, but it clearly gives us *one important kind* of L-freedom – a kind that libertarians can hang their hats on and that's worth wanting and arguing for and so on. After all, in this scenario, the event that settles which option is chosen *is* the conscious decision – i.e., it's the event with a me-consciously-choosing-now phenomenology.²¹

Now, you might think that this misses the point of Pereboom's argument; for you might think that his claim is that the sort of L-freedom that I'm talking about here is insufficient for moral responsibility and basic desert.²² But insofar as moral responsibility and desert are interdefinable with freedom, it should be clear that the sort of L-freedom that I've characterized here is definitely sufficient for a certain *kind* of moral responsibility and desert. We can call them L-responsibility and L-desert, where these are defined in a way that makes it analytic that the sort of freedom they require is the sort of L-freedom that I've been describing here. Now, it is of course true that these are not the only kinds of moral responsibility and desert that one might care about; in particular, we can introduce a notion of *causal-responsibility* and stipulate that it's a kind of moral responsibility that requires causal-control. Moreover, it may be that causal-responsibility is worth wanting and that we don't have it. But so what? That wouldn't change the fact that if we have the sort of L-freedom that I've been talking about, then we also have a libertarian sort of moral responsibility, namely, L-responsibility. Now, you might think that the worry here – the worry that Pereboom is getting at – is that causal-responsibility is *real* moral responsibility. In other words, you might worry that the sort of responsibility that we get from L-freedom – namely, L-responsibility – isn't *real* moral responsibility at all. But this takes us back to the point I was making in Section 2. On my view, the facts that determine which of the various kinds of responsibility count as "real" moral responsibility are facts about the *folk*; in particular, the kind(s) of responsibility that count as real moral responsibility are just the one(s) that the folk have in mind when they think and talk about moral responsibility. If this metaphilosophical view of conceptual-analysis questions is right – and as I pointed out in Section 2, I'm aware that it requires argument – then the question of whether L-responsibility is real moral responsibility is considerably less interesting and important than it might at first seem. Indeed, it's just an empirical question about what the folk mean by 'moral responsibility.'²³ Now, I admit that if L-responsibility wasn't worth wanting at all, then that would be a problem; but it seems to me that the considerations that suggest that L-freedom is worth wanting suggest that L-responsibility is worth wanting as well.

You might think that Pereboom's point is not that causal-responsibility is "real" responsibility, but rather that this is the sort of responsibility that *justifies punishment*.²⁴ But the same kinds of points can be made here all over again. It's obvious that (a) causal-responsibility *causal-justifies* punishment and (b) L-responsibility *L-justifies* punishment but doesn't causal-justify punishment. This leaves open whether L-responsibility *really* justifies punishment, but once again, this is just a question about what the term 'justifies' means in ordinary moral discourse. There is no way to pop out of this circle. *All* of our moral terms can be understood in ways that fit with the kind of L-freedom that I'm talking about – i.e., the kind that we get from TDW-indeterminism – and, of course, these terms can also be understood in ways that fit with agent-causal freedom. Now, my own view is that causal-responsibility and causal-justification are *not* "real" moral responsibility and justification; in other words, I don't think it's built into our ordinary moral practices that agent-causal freedom is required for moral responsibility and the moral justification of punishment. But the point I'm making here is that this doesn't matter; for insofar as L-responsibility and so on are worth wanting, the kind of L-freedom that I'm describing – the kind that we get from TDW-indeterminism – is a worthwhile kind of libertarian free will.

5.6 An Alternate Version of the Luck Argument

As I pointed out above, there are many different versions of the luck argument. In this section, I'll say a few words about one more of these arguments, namely, the following:

The Rollback Version of the Luck Argument: Let's return to Ralph's decision, and let's imagine that God "rolls back" the universe and "replays" the decision 100 times. If the decision is TDW-undetermined, then we should expect that Ralph would choose chocolate pie about 50 times and apple pie about 50 times. But given this—given that Ralph would choose differently in different "plays" of the decision, without anything about his psychology changing—it's hard to see how we can maintain that Ralph authored and controlled the decision. It seems to be a matter of *chance* or *luck* what he chose, and to the extent that this is true, it seems that Ralph didn't author or control the choice.²⁵

The main point I want to make here is that it doesn't follow from the fact that Ralph would choose differently in different "plays" of the decision that he didn't author or control the decision. There is no inconsistency in claiming that (a) Ralph chooses differently in different plays of the decision, and (b) in each of the different plays of the decision, it's *Ralph* who does the choosing, and who authors and controls the choice. Indeed, given that Ralph is making a *torn* decision, the hypothesis that it's *him* who's making the decision (and who's authoring and controlling the decision) seems to *predict* that he would choose differently in different plays of the decision – that seems to be exactly what he would do if he was torn.

6 Responding to the Argument from Naturalism

The purpose of the above discussion was to respond to the luck argument against libertarianism. But it also gives us a response to the argument from spookiness. I have argued here that if our torn decision are neural events, and if they're TDW-undetermined, then they're also L-free. This shows that libertarianism is perfectly consistent with mind-brain materialism and, in particular, with the thesis that human decisions are neural events.

7 Responding to the Argument from Determinism

I now want to discuss the argument from determinism. The idea behind this argument is that libertarianism is false because the kind of indeterminism that it requires is false. In other words, the idea is to argue against libertarianism by arguing against TDW-indeterminism. (TDW-indeterminism is the sort of indeterminism that's needed for *my* version of libertarianism. Other libertarians might want to quibble with this in certain ways, but any (centered) version of libertarianism is going to require something very much like TDW-indeterminism.)

There are numerous ways in which one might try to argue against TDW-indeterminism, but the only really plausible strategy here is to zero in on our torn decisions and attempt to argue that *those very events* aren't undetermined in the manner of TDW-indeterminism. In other words, there's no plausible way to argue for a more general kind of determinism. This is because (a) we currently have no good reason to believe that *all* events are determined (because we have no good reason to believe that all *quantum* events are determined²⁶); and more specifically, (b) we currently have no good reason to believe that all *neural* events are determined (current neuroscientific theory is probabilistic, and there's no good reason to think that this is just a simplification – i.e., for all we know right now, it could be that there are genuine, objective indeterminacies in human neural processes). Given this, it seems that in order to mount a cogent argument against TDW-indeterminism, one would have to focus in on torn decisions in particular and argue that *they* aren't TDW-undetermined.

Another important point to note here is that since libertarians aren't committed to the claim that *all* of our torn decisions are TDW-undetermined (since they claim only that *some* of these decisions are TDW-undetermined), we can't undermine the view by appealing to general arguments from psychology that suggest that our decisions are often influenced by subconscious factors. Libertarians can (and should) admit this. All they need to claim is that *some* of our torn decisions aren't causally influenced by such factors (at the moment of choice).

But even granting all of these points, there are still ways in which one might try to argue that our torn decisions aren't TDW-undetermined. In particular, since our question here is an empirical one, one might try to find some scientific studies that give us reason to doubt that our torn decisions are TDW-undetermined. The most promising arguments here are based on the work of Libet et al. (1983), Tegmark (2000), and Haynes (2011).²⁷ I've responded to all three of these arguments elsewhere (2010, 2014a), but in this paper I'll just say a few words about the one that I think is the hardest to deal with, namely, the argument based on Haynes's study.

Haynes's study can be summarized very quickly: (i) Haynes gave his subjects two buttons, one for the left hand and one for the right; (ii) he told them to make a decision at some point as to which button to push; (iii) he used a very simple method to estimate the time at which the conscious decision occurred; (iv) using fMRI, he found unconscious neural activity in two different regions of the brain that predicted whether subjects would press the left button or the right; and, stunningly, (v) he found that this activity arose *7–10 seconds before the person's conscious decision to push the given button*.

These results generate a serious objection to TDW-indeterminism and libertarianism. We can put the objection like this:

Since neuroscientists could predict which buttons Haynes's subjects would push 7–10 seconds before they consciously chose, their decisions weren't TDW-undetermined; for TDW-indeterminacy requires decisions to be undetermined *at the moment of conscious choice*, and Haynes's findings suggest that his subjects' decisions were already determined 7 seconds before they made their conscious decisions.

One way to respond to this argument would be to claim that the subjects' decisions weren't really *torn* decisions at all – that they were more like Buridan's-ass decisions – and that, because of this, Haynes's results don't tell us anything about torn decisions. But this seems pretty uncharitable to me, and I don't want to respond in this way. Instead, I'd like to grant for the sake of argument that Haynes's subjects' decisions were close enough to torn decisions – or, perhaps better, I'd like to just pretend for the purposes of this discussion that these decisions *were* torn decisions – and I'd like to argue that, even if we grant this point, the above argument doesn't give us any good reason to doubt TDW-indeterminism or libertarianism. In particular, I'd like to argue that when we take note of some of the details of Haynes's study, the argument against TDW-indeterminism falls apart.

There are two details of Haynes's study that I want to focus on. First, the pre-choice brain activity that Haynes found was actually not very good at predicting the outcomes of his subjects' choices. Indeed, it was only 10% more accurate than blind guessing. If we just guess which button subjects are going to push, we'll be right about 50% of the time, whereas if we use information about the brain activity that Haynes found, we'll be right at best 60% of the time. Now, this is definitely statistically significant, so it shows *something*. But it's not clear *what* it shows, and as I'll presently explain, it *doesn't* show (or, indeed, give us any good reason to believe) that TDW-indeterminism is false.

The second important detail of Haynes's study has to do with the specific *regions* of the brain where the pre-conscious-choice neural activity was found; it was found in the *parietal cortex* (or the PC) and *Brodmann area 10* (or BA10). The reason this is important is that these regions aren't associated with free decisions. Rather, they're associated with *plans*, or *intentions*. In particular, they're associated with the generation and storage of plans.²⁸

I'm going to argue that when we put these two facts together – i.e., the facts described in the preceding two paragraphs – they suggest an explanation of Haynes's results that's perfectly consistent with TDW-indeterminism. I'll say in a moment what this explanation is, but first I want to make a background point.

When someone asks you *not* to think about something, it suddenly becomes very difficult to obey them. E.g., if I don't want you to think about chess right now, then one of the worst things I could do is to *tell* you not to think about chess. As soon as I say, "Don't think about chess," it will become very hard for you to avoid thinking about it, even if you sincerely want to obey me.

The same goes for *decisions*. Suppose I say this to you: "In a minute, I'm going to ask you to choose Door 1 or Door 2; but don't do it yet." It's actually somewhat difficult to refrain from thinking of one of the two options in situations like this. I'm not saying you *can't* do this; of course you can; but if we give these instructions to a group of people, it seems very likely that at least *some* of them won't succeed in refraining from thinking of one of the two options. Moreover, if you do think of one of the two options before you're supposed to choose, then this could influence how you end up choosing. Even if you tell yourself not to let this happen – even if you try not to let your pre-choice thoughts influence your decision – you might not succeed, and your pre-choice thoughts might have a causal influence over how you choose.

Now, here's the really important point for us: your choice could be causally influenced by prior goings on in your mind *even if you don't realize this*. You might subconsciously think of Door 1, and you might subconsciously store the plan to pick that door when the time comes. This shouldn't be controversial at all. For here are two things that we *know* to be true of humans: first, it's somewhat difficult for us to avoid thinking about things when people tells us not to think about them; and second, we do all sorts of things unconsciously. We don't do *everything* unconsciously, but we clearly do a *lot* of things unconsciously. When we put these two points together, we get the following (highly probable) hypothesis:

If you tell a group of human subjects that in 60 seconds they're going to be asked to pick Door 1 or Door 2, and if you tell them not to pick yet—in other words, if you tell them to wait until the 60 seconds are up before they choose—then at least *some* of these subjects will (without realizing it) subconsciously think of one of the two doors before the 60 seconds have elapsed, and they'll subconsciously store the plan to pick that door when the time comes.

Again, given what we know about ourselves, this seems extremely plausible. Indeed, it seems to me that it would be very surprising if this weren't true.

These remarks suggest an explanation of Haynes's results that's perfectly consistent with TDW-indeterminism. The explanation I have in mind can be put in the following way:

An explanation of Haynes's results that's perfectly consistent with TDW-indeterminism and libertarianism: A significant percentage of the subjects in Haynes's study (say, 20% of them) unconsciously failed to make truly spontaneous decisions about whether to press the right button or the left button. They genuinely *wanted* to follow Haynes's instructions, but for whatever reason, and without realizing it, they unconsciously formed prior-to-choice plans to push one of the two buttons. They unconsciously stored this information in their brains, and then when the time came, these

plans were activated. In other words, the regions of the brain where these plans were stored were activated. And this brain activity caused the subjects to choose in the ways in which they had unconsciously planned to choose. This explains various features of Haynes's findings. First, it explains why (in some subjects) there was relevant prior-to-choice brain activity in the PC and BA10 regions of the brain—because those regions are associated with the formation and storage of plans. Second, it explains why this brain activity predicted whether subjects would choose to push the left button or the right button. Third, it gives us a nice story about why the predictive brain activity occurred so long before the conscious choice. (We know that humans can make decisions in way less than 7 seconds (see Trevena and Miller (2010) for evidence that we can make decisions in *less than half a second*), and so the fact that Haynes found predictive brain activity 7-10 seconds prior to choice cries out for explanation. But the present explanation gives us a story to tell about this: it's because the relevant brain activity isn't a part of the conscious decision at all—it's associated with something completely distinct from, and prior to, the decision.) Fourth and finally, the present explanation of Haynes's data explains why the brain activity in the PC and BA10 is only 10% better, vis-à-vis accurate prediction, than blind guessing. The reason is that *not all* subjects unconsciously formed plans about what they were going to do. Only *some* of them did. Most of the subjects managed to avoid doing this, and so most of them succeeded in making truly spontaneous decisions. (Of course, if libertarians endorse this explanation, their claim won't be that most of us are L-free but some of us aren't; their claim will be that *all* of us *sometimes* fail to be L-free. The idea here is that we're all *sometimes* driven by things like unconscious plans; but we aren't *always* driven by such things.)

The first point to note about this explanation of Haynes's results is that if it's right, then there's no problem here for TDW-indeterminism or libertarianism. All Haynes's results show, if my explanation is correct, is that some of our decisions are influenced by unconscious factors. But we already *knew* this. Libertarians don't think that *all* of our torn decisions are TDW-undetermined. Moreover, they can admit (and *should* admit) that many of our torn decisions are causally influenced by unconscious factors in freedom-undermining ways – and in ways that make it the case that they're not TDW-undetermined or L-free. What libertarians claim is that this isn't *always* the case; again, their claim is merely that *some* of our torn decisions are TDW-undetermined. But given this, if my explanation of Haynes's findings is correct, then those findings don't give us any good reason to doubt libertarianism because they don't give us any good reason to think *none* of our torn decisions is TDW-undetermined. All these findings show is that *some* of our torn decisions (or more precisely, some decisions that are similar to torn decisions in certain ways) aren't TDW-undetermined – and this is perfectly consistent with the libertarian view that some of our torn decisions *are* TDW-undetermined.

In order to block my argument here – in order to maintain that Haynes's results create a problem for TDW-indeterminism – you'd have to endorse the following view:

The anti-libertarian interpretation of Haynes's data: Haynes's data suggest that TDW-indeterminism is false, because they suggest that *none* of our torn decisions is TDW-undetermined, because they suggest that *all* of our decisions (or at least all of our torn decisions) are determined (or at least significantly causally influenced) by prior-to-conscious-choice neural activity of the kind that Haynes found.

The first point I want to make about this view is that there's simply no good reason to believe it – Haynes's data just don't support this sweeping conclusion. And from this alone, it already follows that Hayne's data don't give us any good reason to doubt TDW-indeterminism.

But I want to argue for a stronger claim. I want to argue that my explanation of Haynes's data is *superior* to the anti-libertarian interpretation of Haynes's data. The reason that my

explanation is superior is that it's more explanatory. As we saw above, my explanation explains all of the following: (i) why we found predictive brain activity in the PC and BA10 regions; (ii) why this brain activity predicted whether subjects would push the left button or the right button; (iii) why this activity occurred so long before the conscious choice; and (iv) why this activity was only 10% better, in terms of accurate prediction, than blind guessing. In contrast, the anti-libertarian interpretation doesn't explain any of these things. Now, advocates of the anti-libertarian interpretation could explain (i)–(iii) by stealing *my* explanation – i.e., by claiming that the brain activity that Haynes found is associated with the formation and storage of unconscious pre-choice plans to push one of the two buttons. Advocates of the anti-libertarian interpretation would just have to add that this happens in *all* subjects – and, indeed, *all torn decisions*. But this doesn't fit at all with (iv); (iv) fits much better with my interpretation. Moreover, when we realize that the best version of the anti-libertarian interpretation is the one that steals the core of my explanation – i.e., the part about the formation of unconscious pre-choice plans – it throws into full relief how weird and unjustified it is for advocates of the anti-libertarian interpretation to claim that this occurs in *all* cases. In other words, it lays bare the fact that there's simply no evidence for this sweeping universal conclusion.

So I don't think we have any good reason to believe the anti-libertarian interpretation of Haynes's data, and I think that my explanation of Haynes's data is superior to the anti-libertarian interpretation.

8 HB-Libertarianism as an Open Empirical Question

The arguments of this paper are suggestive of the conclusion that, as of right now, we don't have any good arguments for or against HB-libertarianism. I obviously haven't done all the work necessary to establish this result here, but I think it's true, and so I think that the question of whether HB-libertarianism is true is an open question. Moreover, I think it's an open *empirical* question. In particular, I think the question of whether HB-libertarianism is true essentially boils down to the empirical question of whether TDW-indeterminism is true. I've partially argued for this claim here because I've argued that if TDW-indeterminism is true, then HB-libertarianism is true as well. But I think it can also be argued that if TDW-indeterminism *isn't* true, then HB-libertarianism isn't true either.²⁹

Notes

- 1 I would like to thank Chris Franklin, Kristin Mickelson, and Leigh Vicens for comments on earlier versions of this paper.
- 2 Some of the philosophers who have defended libertarian views of one kind or another are: Bramhall (1655), Reid (1788), Chisholm (1964), Taylor (1966), Campbell (1967), Wiggins (1973), Thorp (1980), Nozick (1981), van Inwagen (1983), Kane (1985, 1996, 1999), Rowe (1987), Donagan (1987), Ginet (1990, 2002), Clarke (1993, 1996), McCall (1994), Goetz (1997), McCann (1998), O'Connor (2000), Ekstrom (2000), myself (2004, 2010), Pink (2004), Griffith (2007), Lowe (2008), Franklin (2011), Mawson (2011), Steward (2012), and Todd (2016).
- 3 In case you're wondering, the 'HB' stands for 'human being' and the 'C' stands for 'conceptual'.
- 4 If we wanted to, we could replace C-libertarianism with the following weaker thesis: "L-freedom is a legitimate kind of free will." If we proceeded in this way, then libertarianism would leave open whether there are other legitimate kinds of free will and, in particular, whether there are

legitimate kinds of free will that are compatible with determinism. I like this way of defining the view, but I won't worry about this here; I'll just assume that libertarianism is supposed to entail incompatibilism.

- 5 If they wanted to, libertarians could define 'appropriately non-random' in a way that implied that virtually Hume-free decisions don't count as non-random. But in order to do this, they would need to more or less build clause (b) of the definition of L-freedom into the definition of appropriate non-randomness—and so it wouldn't change anything in a substantive way. I think it's simpler and clearer to be less strict about what's needed for appropriate non-randomness and then to say that L-freedom requires that the indeterminacy generates (or procures, or enhances, or whatever) the non-randomness.
- 6 I argue for both of these claims in my (2010). Claim (b) might seem surprising. The argument for it is based on an argument for the claim that in this scenario, the answer to the conceptual-analysis question would be that the concept *free will* is identical to the concept *the freedom-like ability that humans actually have, whatever that turns out to be*; the answer would *not* be the specific freedom-like ability that humans happen to have.
- 7 I give much more thorough arguments for claim (a) in Chapter 8 of (Balaguer 2021) and for claim (b) in Chapter 2 of (Balaguer 2010).
- 8 Here's an analogy. The concept *even prime greater than 2* is obviously incoherent; but this doesn't mean that there's no such thing as the concept of an even prime greater than 2, and in fact, there clearly *is* such a concept. If there were no such concept, then the sentence 'There's no such thing as an even prime greater than 2' would be *gibberish*; but it's *not* gibberish; indeed, it's *true*.
- 9 We can also argue against the importance of the C-question by arguing that multiple concepts of freedom are at work in ordinary discourse so that there's no unique determinate answer to the what-is-free-will question. I argued for this claim in my (Balaguer 2010), Section 2.7.
- 10 Arguments of this general kind trace back at least to Kant. For a more recent version, see van Inwagen (1983). And for more on this general kind of argument, see Chan (2017).
- 11 See, e.g., Hobbes (1651), Hume (1748), Hobart (1934), Waller (1988), Double (1991), Bernstein (1995), Clarke (1995, 2002), Fischer (1999), Mele (1999a, 1999b), Haji (1999), O'Connor (2000), G. Strawson (2000), Berofsky (2000), van Inwagen (2002), and Levy (2011).
- 12 Just about everyone who has defended libertarianism has responded to arguments of this general kind. To name just a few who have responded to arguments of the kind outlined in the text, see, e.g., Kane (1999), Ginet (2007), Balaguer (2010), and Franklin (2011).
- 13 Agent-causal views have been developed by Reid (1788), Chisholm (1964), Taylor (1966), Thorp (1980), Rowe (1987), Donagan (1987), Clarke (1993), O'Connor (2000), Griffith (2007), Mawson (2011), and Steward (2012).
- 14 Event-causal views have been developed by Wiggins (1973), Nozick (1981), van Inwagen (1983), Kane (1985, 1996), Ekstrom (2000), Balaguer (2004, 2010), and Franklin (2011).
- 15 Non-causal views have been developed by Ginet (1990, 2002), McCall (1994), Goetz (1997), McCann (1998), Pink (2004), and Lowe (2008).
- 16 On some ways of conceiving of event-causal libertarianism, it has non-causal elements in it. For instance, many event-causal libertarians would say that our L-free decisions are probabilistically caused by our reasons for choosing; but one might think that what this means is that our reasons for choosing deterministically cause there to be certain objective probabilities of the various options being chosen and that, beyond this, which option ends up being chosen is uncaused. In other words, you might think that probabilistic event causation is just a mix of deterministic event causation and non-causation. Event-causalists don't have to endorse this line, but it strikes me as plausible.
- 17 Some people interpret Ekstrom (2000) as endorsing deliberativism; I think this is a misinterpretation, and in private correspondence, she has told me that she agrees with my interpretation of her view as a centered one.
- 18 I should say that it's *possible* to make a torn decision while in a Buridan's-ass situation – because you could be weird enough to *care* which can of Campbell's tomato soup you get, and so you could

- feel genuinely torn about it. But *most* of us don't make torn decisions in Buridan situations. For example, in the above situation, most of us would just grab a can of soup without thinking about it.
- 19 On some ways of defining 'determined,' expressions like 'partially determined' don't make sense. But I'm not using the term in any such way here.
 - 20 In saying this, I don't mean to commit to a type-type identity theory according to which the mental kind *decision* is identical to some physical kind; I just mean to commit to the token-token view that any specific decision is a neural event.
 - 21 For more on why this sort of L-freedom is worth wanting, see Balaguer (2010).
 - 22 Chris Franklin raised this worry to me.
 - 23 For more on these issues, see Balaguer (2020) and Chapters 8 and 19 of Balaguer (2021).
 - 24 Leigh Vicens raised this worry to me.
 - 25 van Inwagen constructs an argument of this kind in his (2002).
 - 26 Some philosophers seem to think that we have good positive reason to think that some quantum events are *undetermined*. This, I think, is false – we have no good reason to believe this claim. But we also have no good reason to disbelieve it. In short, the question of whether there are any quantum events that are genuinely undetermined is an open empirical question.
 - 27 Of course, these aren't the only scientific works that one might use to argue against TDW-indeterminism. Others include Festinger (1957), Milgram (1969), Isen and Levin (1972), and Velmans (1991). For a survey of potentially relevant scientific works, see Wegner (2002).
 - 28 For evidence that BA10 is associated with the storage of plans and intentions, see, e.g., Burgess, Quayle, and Frith (2001) and Haynes, Sakai, Rees, et al. (2007). And for evidence that the PC is associated with the generation of plans, see, e.g., Desmurget and Sirigu (2009).
 - 29 I argued for this claim in Balaguer (2010). I did this by arguing for the following two claims: (i) if our torn decisions aren't TDW-undetermined, then they're not (fully) L-free; and (ii) if our torn decisions aren't TDW-undetermined – and, hence, aren't (fully) L-free – then it's very likely that decisions of various other kinds that we might have thought were L-free (e.g., our leaning decisions and Buridan's-ass decisions) aren't (fully) L-free either. Claim (i) follows pretty easily from some things that I've argued for in this paper, but I haven't said anything in this paper about claim (ii).

Bibliography

- Balaguer, M. (2004). A coherent, naturalistic, and plausible formulation of libertarian free will. *Nous* 38: 379–406.
- Balaguer, M. (2010). *Free Will As An Open Scientific Problem*. Cambridge, MA: MIT Press.
- Balaguer, M. (2014a). *Free Will*. Cambridge, MA: MIT Press.
- Balaguer, M. (2014b). Replies to McKenna, Pereboom, and Kane. *Philosophical Studies* 169: 71–92.
- Balaguer, M. (2020). Moral folkism and the deflation of (lots of) normative and metaethics. In: *Abstract Objects: For and Against* (ed. J. Falguera Lopez and C. Martinez-Vidal), 297–318. Synthese Library.
- Balaguer, M. (2021). *Metaphysics, Sophistry, and Illusion*. Oxford: Oxford University Press.
- Bernstein, M. (1995). Kanean libertarianism. *Southwest Philosophy Review* 11: 151–157.
- Berofsky, B. (2000). Ultimate responsibility in a determined world. *Philosophy and Phenomenological Research* 60: 135–140.
- Bramhall, J. (1655). *A Defense of True Liberty*. Reprinted 1977. New York: Garland Pub.
- Burgess, P.W., Quayle, A., and Frith, C.D. (2001). Brain regions involved in prospective memory as determined by positron emission tomography. *Neuropsychologia* 39: 545–555.
- Campbell, C.A. (1967). *In Defense of Free Will*. London: Allen & Unwin.
- Chan, R. (2017) *Making metaphysics matter: essays on reasons and person*, PhD dissertation, University of Notre Dame.
- Chisholm, R. (1964). Human freedom and the self. Reprinted in: *Free Will* (ed. G. Watson), 1982, 24–35. Oxford: Oxford University Press.
- Clarke, R. (1993). Toward a credible agent-causal account of free will. *Nous* 27: 191–203.

- Clarke, R. (1995). Indeterminism and control. *American Philosophical Quarterly* 32: 125–138.
- Clarke, R. (1996). Agent causation and event causation in the production of free action. *Philosophical Topics* 24: 19–48.
- Clarke, R. (2002). Libertarian views: critical survey of noncausal and event-causal accounts of free agency. In: Kane (2002), 356–385.
- Dennett, D. (1978). On giving libertarians what they say they want. In: *Brainstorms*. Cambridge, MA: MIT Press.
- Desmurget, M. and Sirigu, A. (2009). A parietal-premotor network for movement intention and motor awareness. *Trends in Cognitive Sciences* 13: 411–419.
- Donagan, A. (1987). *Choice*. London: Routledge & Kegan Paul.
- Double, R. (1991). *The Non-Reality of Free Will*. New York: Oxford University Press.
- Ekstrom, L.W. (2000). *Free Will: A Philosophical Study*. Boulder, CO: Westview Press.
- Festinger, L., (1957). *A Theory of Cognitive Dissonance*. Palo Alto: Stanford University Press.
- Fischer, J.M. (1999). Recent work on moral responsibility. *Ethics* 110: 93–139.
- Franklin, C. (2011). Farewell to the luck (and mind) argument. *Philosophical Studies* 156: 199–230.
- Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.
- Ginet, C. (2002). Reasons explanations of action: causalist versus noncausalist accounts. In: Kane (2002), 386–405.
- Ginet, C. (2007). An action can be both uncaused and up to the agent. In: *Intentionality, Deliberation, and Autonomy* (ed. C. Lumer and S. Nannini), 243–256. London: Routledge.
- Goetz, S. (1997). Libertarian choice. *Faith and Philosophy* 14: 195–211.
- Griffith, M. (2007). Freedom and trying: understanding agent-causal exertions. *Acta Analytica* 22: 16–28.
- Haji, I. (1999). Indeterminism and Frankfurt-style examples. *Philosophical Explorations* 2: 42–58.
- Haynes, J.D. (2011). Decoding and predicting intentions. *Annals of the New York Academy of Sciences* 1224: 9–21.
- Haynes, J.D., Sakai, K., Rees, G. et al. (2007). Reading hidden intentions in the human brain. *Current Biology* 17: 323–328.
- Hobart, R.E. (1934). Free will as involving determinism and inconceivable without it. *Mind* 43: 1–27.
- Hobbes, T. (1651). *Leviathan*. Reprinted 1962. New York: Collier Books.
- Hume, D. (1748). *An Inquiry Concerning Human Understanding*. Reprinted 1955. Indianapolis, IN: Bobbs-Merrill.
- Isen, A. and Levin, P. (1972). Effect of feeling good on helping. *Journal of Personality and Social Psychology* 21: 384–388.
- Kane, R. (1985). *Free Will and Values*. Albany, NY: State University of New York Press.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.
- Kane, R. (1999). Responsibility, luck, and chance: reflections on free will and indeterminism. *Journal of Philosophy* 96: 217–240.
- Kane, R. (ed.) (2002). *The Oxford Handbook of Free Will*. New York: Oxford University Press.
- Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.
- Libet, B., Gleason, C., Wright, E., and Pearl, D. (1983). Time of conscious intention to act in relation to cerebral potential. *Brain* 106: 623–642.
- Lowe, E.J. (2008). *Personal Agency: The Metaphysics of Mind and Action*. Oxford: Oxford University Press.
- Mawson, T.J. (2011). *Free Will: A Guide for the Perplexed*. New York: Continuum Press.
- McCall, S. (1994). *A Model of the Universe*. Oxford: Oxford University Press.
- McCann, H. (1998). *The Works of Agency: On Human Action, Will, and Freedom*. Ithaca, NY: Cornell University Press.
- Mele, A. (1995). *Autonomous Agents*. New York: Oxford University Press.
- Mele, A. (1999a). Kane, luck and the significance of free will. *Philosophical Explorations* 2: 96–104.
- Mele, A. (1999b). Ultimate responsibility and dumb luck. *Social Philosophy and Policy* 16: 274–293.
- Milgram, S. (1969). *Obedience to Authority*. New York: Harper and Row.

- Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA: Harvard University Press.
- O'Connor, T. (2000). *Persons and Causes*. New York: Oxford University Press.
- Pereboom, D. (2014). The disappearing agent objection to event-causal libertarianism. *Philosophical Studies* 169: 59–69.
- Pink, T. (2004). *Free Will: A Very Short Introduction*. Oxford: Oxford University Press.
- Reid, T. (1788). *Essays on the Active Powers of the Human Mind*. Reprinted 1969. Cambridge, MA: MIT Press.
- Rowe, W. (1987). Two concepts of freedom. *Proceedings of the American Philosophical Association* 62: 43–64.
- Steward, H. (2012). *A Metaphysics for Freedom*. Oxford: Oxford University Press.
- Strawson, G. (2000). The unhelpfulness of indeterminism. *Philosophy and Phenomenological Research* 60: 149–155.
- Taylor, R. (1966). *Action and Purpose*. Englewood Cliffs, NJ: Prentice-Hall.
- Tegmark, M. (2000). The importance of quantum decoherence in brain processes. *Physical Review E* 61: 4194.
- Thorp, J. (1980). *Free Will*. London: Routledge & Kegan Paul.
- Todd, P. (2016). Strawson, moral responsibility, and the “order of explanation”: an intervention. *Ethics* 127: 208–240.
- Trevena, J. and Miller, J. (2010). Brain preparation before a voluntary action: evidence against unconscious movement initiation. *Consciousness and Cognition* 19: 447–456.
- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.
- van Inwagen, P. (2002). Free will remains a mystery. In: Kane, (2002), 158–177.
- Velmans, M. (1991). Is human information processing conscious? *Behavioral and Brain Sciences* 14: 651–669.
- Waller, B. (1988). Free will gone out of control: a review of R. Kane's *Free Will and Values*. *Behaviorism* 16: 149–162.
- Wegner, D.M. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.
- Wiggins, D. (1973). Towards a reasonable libertarianism. Reprinted in: *Essays on Freedom of Action* (ed. T. Honderich), 31–61. London: Routledge & Kegan Paul.

Part IV

Moral Responsibility

Children and Moral Responsibility¹

MEGHAN GRIFFITH

Are children morally responsible for their choices and actions? If so, are there any important differences between their responsibility and ours? I think the answer to both of these questions is a qualified “yes.” The challenge is determining exactly how these yeses should be qualified. When are children responsible and when are they not? If they are responsible, in what ways might their responsibility differ from ours? If we are not careful when answering these questions, we run the risk of erring in one direction or the other. We will either disrespect the personhood and moral agency of children by underestimating their responsibility, or we will be overly morally demanding and assume that they are, for example, as blameworthy as we are.² The truth, I think, is more complex and nuanced.

Philosophers are no strangers to the complex and nuanced, but children are not often discussed in the moral responsibility and free will literature. Maybe it isn’t clear what philosophical wisdom we would gain. Or perhaps the questions about children are regarded as intriguing, but not central to our main concerns. Understanding responsibility in children, however, could very well be central to our main concerns. In the first place, part of the value of studying free will and responsibility is its potential to influence how we relate to one another.³ Most of us have important relationships with children. It would seem odd, then, to ignore important philosophical questions about their agency and personhood. Studying responsibility in children could help us reflect on how we ought to treat persons who are significant to us. But this may not be the only benefit. Studying children’s responsibility might illuminate the kinds of capacities and factors that play a role in responsible agency. In other words, studying children’s responsibility might not just teach us about children. It might teach us about responsibility.⁴

1 Kinds of Deficiency

In this chapter, I examine the issue of children’s responsibility from three different, but not incompatible, angles. First, I look at standard threshold conditions for responsibility as developed in the free will and responsibility literature. I provide considerations to guide us in

deciding whether or when children can meet them. Second, I discuss David Shoemaker's tripartite view of responsibility according to which an agent might be responsible along one or two dimensions without being responsible along all three. Third, I address the idea of degrees of blameworthiness. I utilize the idea that a decrease in the degree of blameworthiness can be tied to an increase in the difficulty one has in doing the right thing (see Faraci and Shoemaker 2010, 2014; Nelkin 2014). There is a lot to explore about why acting rightly could be more difficult for children. But I discuss one particular developmental phenomenon with wide-ranging explanatory potential. The phenomenon in question is the development of narrative capacities. In particular, the late developing capacity to understand one's own life and identity as a narrative might provide important clues. If my speculations about narrative are on the right track, this could tell us something important about responsible agency in general.

In the discussion that follows, I draw on different prominent and influential responsibility schemas without defending any one in particular and without settling various disputes (For example, I appeal to both reasons-responsive views and identificationist views without adjudicating between them. I do not worry about incompatibilism versus compatibilism or the debate between leeway and sourcehood. I do not defend responsibility pluralism against monism or vice versa.). My aim is to consider children's responsibility from a number of angles, and I recognize that my examination does not include all the many worthy views of responsibility.

2 Threshold Conditions

2.1 *Knowledge of Consequences*

Moral responsibility is typically characterized as having a control, or freedom-relevant, condition, and an epistemic condition. Although these conditions are often applied to particular cases in order to assess responsibility for a given action or choice, we can also use them as a way of distinguishing appropriate candidates for responsibility from those who do not (yet?) meet the threshold to be responsible agents. Thus, a person who is not yet capable of controlling himself in the way required or of acquiring the relevant knowledge will not be part of the "responsibility community."⁵

I begin with the epistemic condition. For our purposes it should suffice to give a rough characterization. Following Aristotle, we can say that if an agent is nonculpably ignorant of the relevant circumstances of her action, including the likely consequences, she cannot be held responsible. We often suppose that children can't acquire the appropriate knowledge of circumstances and consequences. Imagine the preschooler who does not anticipate, for example, that his description of someone's appearance will lead to hurt feelings; or consider the even younger child who does not anticipate that his throwing toys will lead to hurt bodies. So we are inclined to excuse on the grounds that they didn't "know better."

But deficiencies in knowledge of circumstances and consequences will not explain all of our practices and intuitions. Anyone who has spent a fair amount of time around children will confirm that children often *do* know that, say, hitting causes pain. And developmental psychology supports this. Psychologists now claim that children understand cause and effect pretty well. Even as early as age 2, children are able to produce causal explanations if they are familiar with the subject matter (Gopnik 2009, p. 35). Very young children also make predictions on the basis of their causal knowledge (Gopnik 2009, p. 38). So while the preschooler and younger child in the above examples may be exempt in those particular cases, we need to

think carefully before assuming that children at young ages can *never* really understand the effects of their behavior. Are these children responsible?

2.2 *Reasons Receptivity*

Whether children who understand causes and consequences are responsible may depend on what else they are able to understand. Understanding consequences is not the same as understanding moral wrongness.⁶ And some children might understand the former but not the latter. But what does responsibility require? On some current accounts of responsibility, the ability to make judgments about *non-moral* reasons is sufficient for responsibility. On these kinds of views, psychopaths, for example, are responsible, even though they cannot be motivated by *moral* reasons.⁷ Does this mean that if a child is able to recognize and respond to various prudential reasons (e.g., negative and positive reinforcements), then she is responsible, even if she cannot recognize moral reasons?⁸ Other prominent views will say no; on these views, no one is responsible until she can recognize moral reasons. On such views, if one fails to be receptive to moral reasons this means that one fails to satisfy the control, or freedom-relevant, condition of responsibility (see McKenna 2017, p. 27). The idea is that someone cannot properly guide and control her own behavior if she cannot both recognize and react to *the right kinds* of reasons, such as moral reasons (see, for example, Fischer and Ravizza 1998, p. 76ff).

With all of this in mind, we might take a closer look at what children are plausibly able to do. It seems fairly clear that children struggle more than adults do with recognizing the moral reasons the world provides. But matters are complicated. Note the following background from developmental psychologists Monisha Pasupathi and Cecilia Wainryb:

Developmentally, children develop a sense of moral concerns as distinct from other types of concerns quite early in life [Nucci 1981; Turiel 1998]. Even very young children view harming others as wrong [Turiel 1998; Smetana 2006], show evidence of empathy with others' experiences [Eisenberg, Carlo, Murphy, and Van Court 1995; Rotenberg and Eisenberg 1997], and consider moral issues as distinct from other domains of social cognition [e.g., Nucci 1981]. However, that is not to say that developmental changes have no implications for the nature of judgments in situations of competing concerns. Children weight competing concerns in mixed situations differently across different developmental periods [e.g., Helwig, 2002; Killen et al., 2002; Komolova & Wainryb, submitted; Shaw & Wainryb, 2006], making developmental changes one of the critical considerations in just what children and adolescents make of complex situations involving harm (Pasupathi and Wainryb 2010, p. 60).⁹

This indicates that children recognize moral reasons earlier than one might generally suppose. Nonetheless, there may be responsibility differences insofar as it may be harder for them to properly weight moral considerations in complex situations (more will be said about this below).

But when we claim that children can distinguish between moral concerns and other concerns, we might wonder whether these children are merely recognizing that certain norms are categorized as moral. Could it be that they recognize a distinction between moral and conventional without these moral concerns providing special reasons for action? In this way, perhaps they are like psychopaths, who have been shown in some studies to be able to do as well as non-psychopathic adults on a certain kind of moral/conventional distinction test.¹⁰ Psychopaths thus in some sense can “see” the moral “reasons” but are not motivated by them as such. Applying this possibility to children is interesting, but a parallel seems unlikely,

given that (non-psychopathic) children appear to have more in common, morally speaking, with non-psychopathic adults than they do with psychopaths insofar as (non-psychopathic) children demonstrate crucial other-regarding capacities at young ages. For example, unlike psychopaths, very young (non-psychopathic) children are troubled by the pain and distress of others (see Gopnik 2009, p. 213). Furthermore, even before age 2, they try to mitigate this distress when they witness it (211). As Allison Gopnik puts it: “Eighteen-month-olds are both empathic and altruistic – they feel the pain of others and they try to ameliorate it” (Gopnik 2009, p. 212). This does not, of course, translate directly into the ability to recognize moral reasons and be motivated by them. But by the time a child can articulate the difference between, for example, prohibitions against harm and conventional prohibitions, why suppose that these empathic and altruistic tendencies fall away? Why not suppose that these tendencies factor into their motivations and judgments just as they do in ours?¹¹

2.3 *Reasons Reactivity*

As we have seen, it is a complicated question whether children are as able to recognize moral reasons as we are. It seems that fairly early on, children have a more sophisticated ability to do so than we might have supposed, though they may have more trouble than we do assigning the proper weights to these reasons. But we know from the responsibility literature that control is not entirely dependent on what reasons one is able to *recognize*. One might recognize the appropriate reasons but be unable to respond to them. We need both receptivity and reactivity.¹² Other motivational elements can hinder reactivity. For the sake of simplicity I will put these elements under the heading of “desires.” Could it be that even when children properly rank reasons, their desires prevent them from acting accordingly?

Let’s look at so-called failures of self-control. These are cases in which someone acts in accordance with competing motivations against her best judgment (Mele 2012, pp. 91–93). We are all – adults and children – prone to such failures. But we think it rare that adult failures are due to *irresistible* desires. Thus, we assume that adults satisfy the control condition and are responsible. Could it be that for children, such desires really are irresistible?

I think this is unlikely. For example, between the ages of 3 and 5, many children become quite successful at delaying gratification, as shown in the famous Mischel experiments (often referred to as the marshmallow experiments) (Gopnik 2009, p. 59).¹³ For example, they can resist the desire for a treat if told that waiting will get them two instead. Now of course the desire for a treat may be different in important ways from the desire to hit when angry. But even with angry outbursts, it seems reasonable to think that such motivations are not literally irresistible. Children are often able to alter their responses to anger depending on the circumstances (e.g., the angry child might hit only when no teacher or parent is looking). So I do not think we can say that children are unable to control themselves. But we might wonder whether it is harder for them to do so. More will be said below about the role of difficulty in responsibility. But for now, we should conclude that children’s desires do not appear to make them *unable* to react to appropriate reasons.

Could desires play an important role in a different way (other than by affecting reactivity)? Perhaps we ought to think about higher order desires. Consider an identificationist (or mesh or real or deep self) theory of free will and responsibility. The most famous version is Harry Frankfurt’s theory. In his landmark paper, “Freedom of the Will and the Concept of a Person,” he claims that free will consists in being able to have the will that one wants. Those who are unable to do so, due to an addiction, say, or some other internal constraint, lack free will. But sometimes a human being lacks free will because he is unable to care about his own will.

He is a wanton who lacks free will “by default” (Frankfurt 1971, p. 15). So could it be that children are not responsible because they are wantons (Frankfurt himself attaches this label to very young children. See Frankfurt 1971, p. 11)? Caring about one’s will seems to require being able to reflect on one’s will. There are indications that young children understand their own minds in a third-personal way before they are able to engage in introspection from the first-person (Bear and Bloom 2017, p. 480). For example, some studies appear to show that children cannot (perhaps until age 7 or so), accurately tell when their own behavior is intentional (Bear and Bloom 2017, pp. 480–481). Perhaps this indicates that they are unable to reflect sufficiently on their own motivations to avoid wantonness. On the other hand, children in this age group do develop “executive control,” which may involve some evaluative capability.¹⁴ Four- and five-year-olds are able to imagine what they will want or how they will feel at some future time (Gopnik 2009, pp. 147–148). As Gopnik suggests, this is what executive control requires: “I don’t just have to imagine alternative ways the world might be, I have to imagine alternative ways that I might be” (Gopnik 2009, p. 148). They can imagine, for example, that they will want the two cookies later, so they can try to avoid eating the one cookie now (as discussed above). And they can engage in strategies like closing their eyes in order to resist their temptation for immediate gratification (Gopnik 2009, p. 148). While this may not involve an explicit evaluation of one’s will, it does seem to involve an implicit one. These children are engaged in a deliberate attempt to change the strength of their own motivations because they recognize on some level that these motivations are problematic.¹⁵

In any case, under the right social circumstances, metacognitive abilities improve throughout development (Duckworth, Gendler and Gross 2014, p. 204). So surely older children and adolescents will typically not be wantons because they will have strengthened metacognitive abilities and will be able to reflect upon their own wills. Then the question becomes whether these older children should be held fully responsible (provided that the other conditions are also satisfied).

3 Free Will Beliefs

Thus far, we have looked at the ability in children to recognize reasons and react to them, as well as the ability to engage in reflection about their own motivations. But we might also wonder about their ability to have accurate beliefs about what they are able to do. We might think that children will not be able to recognize and react to the right reasons if they do not believe themselves capable of acting in various ways. Recent experiments in developmental psychology examine free will beliefs in children. In one series of five experiments undertaken by Tamar Kushnir, Alison Gopnik, Nadia Chernyak, Elizabeth Seiver, and Henry M. Wellman, the intuitions of 4 and 6-year-olds were tested in a range of cases. I will discuss two of these experiments.

In one experiment in the series, 4-year-olds were asked in a *Free Drawing* trial to draw a dot. Then, in a *Constrained Drawing* trial, they were asked to draw a line, but their hands were held so as to force them to draw a dot (Kushnir et al. 2015, p. 84). After each, the children were asked “could you have drawn the line?” Most of the children (71%) said “yes” after the *Free Drawing* trial, while a minority (19%) said “yes” after the *Constrained Drawing* trial (Kushnir et al. 2015, p. 85). In another experiment, 4 and 6-year-olds were asked to name foods they did and did not like and were asked corresponding questions about whether they could choose to eat what they didn’t like or choose not to eat what they did like (Kushnir et al. 2015, p. 94).¹⁶ The study authors sum up the results of this particular experiment as follows:

4-year-olds, but not 6-year-olds, mostly said that they were *not* free to act against their own stated desires. Together with the results of the previous analysis, this suggests that 4-year-olds feel constrained by their own stated motivations, and even that they may treat them as similar to physical impossibility. (Kushnir et al. 2015, p. 95)

And the results of this series of experiments purportedly

reveal that 4-year-old children have intuitions about their own and others ability to do otherwise which in some ways resemble our adult intuitions, and in some ways are remarkably different. First, 4-year-old children distinguish between agents' free actions and actions that are physically and mentally constrained. We also find, however, that is [sic] not until age 6 that children consistently reason that psychological motivations do not necessarily constrain our actions, and that we have the ability to choose alternative actions that go against our desires. (Kushnir et al. 2015, p. 98)

Other researchers have also found that young children's intuitions about free will are different from adults'. In a study by J. Phillips and P. Bloom, children ages 4 to 7 were asked whether certain acts were possible or impossible. The youngest participants did not distinguish between fantastical acts (e.g., "a boy throwing his hat into the air and the hat transforming into a candy bar") and immoral ones (e.g., "a boy stealing a candy bar"), saying both were impossible (Bear and Bloom 2017, p. 479). In a second study, participants were asked which scenarios were magical. Again, the youngest subjects grouped physical impossibility with moral impermissibility, saying both were magical (these findings – not yet published – are discussed in Bear and Bloom 2017, p. 479).¹⁷

What does all of this tell us about the development of responsibility? On the one hand, the beliefs children have about inability (to resist their desires, to act immorally) clearly do not translate into actual inability to perform the kinds of actions involved. As Bear and Bloom note, children regard immoral actions as impossible "*even though* they have observed such actions and have engaged in them themselves" (Bear and Bloom 2017, p. 479. My emphasis). Similarly, 4-year-olds do seem to be able to resist their desires, as discussed above.

But, on the other hand, in both cases, there appears to be a failure by these children to understand their own motivational and choosing processes. And such failures of self-knowledge have been said to undermine responsibility.¹⁸ The question, then, is not so much about what kinds of actions children can or cannot perform, but whether in performing such actions, they are able to satisfy responsibility conditions such as reasons-responsiveness. This is not an easy question to answer. A case can be made on either side. In favor of satisfying reasons-responsiveness, one could argue that were the children to have sufficient reason to do otherwise, they would do so. On the other hand, we might wonder whether this is sufficient for responsibility because these 4-year-olds do not see their reasons *as reasons* but as constraints (akin to having their hands held in place). One might suggest that this leads to a limited ability to evaluate or weight reasons in the appropriate way. How can the child properly evaluate what she has most reason to do if she cannot even know that her desire is just one reason among the other possible competing reasons? Note also that on John Martin Fischer and Mark Ravizza's influential reasons-responsiveness account, in order for there to be responsibility, one must have "taken responsibility" for the mechanism from which one acts and this requires coming to understand how one's agency operates in the world.¹⁹ Children who are unaware of the way their choosing processes function would not be able to take responsibility for these mechanisms. But on the other side, some theorists claim that one need

not be consciously aware of one's reasons as reasons in order to have sufficient control for moral responsibility.²⁰ What we decide about these 4-year-olds, then, may depend on what we think is required for reasons-responsiveness.

Interestingly, the authors of the series of experiments about free will beliefs (Kushnir et al) speculate that as children begin to have free will beliefs more like ours, where desire is separated from will or choice, their control over their behavior increases as a result. Preliminary evidence comes from the delayed gratification experiments. Self-control increases when children understand that they can intervene in their own mental states (Kushnir et al. 2015, p. 98).

4 Kinds of Responsibility

Another way we might think about children and responsibility is in terms of different kinds of responsibility, such that children are responsible along some dimensions but not others. David Shoemaker argues forcefully and eloquently for a tripartite view of responsibility whereby certain "marginal agents" may be responsible in some of the senses but not all of them. Our ambivalence about such agents is explained by the idea that they are both responsible and not responsible, with respect to different dimensions. He has in mind agents with various disorders or disabilities, such as those with dementia or psychopathy. But could we apply this to children in order to explain our ambivalence and clarify the grounds of our responses? It is an intriguing question.

Building on the influential work of P.F. Strawson, Shoemaker takes a quality of will approach to moral responsibility. Strawson suggests that responsibility is about our "reactive attitudes" which "are essentially reactions to the *quality of others' wills* towards us, as manifested in their behavior; to their good or ill will or indifference or lack of concern" (Strawson 2003, p. 83) (as quoted in Shoemaker 2015, p. 7; his emphasis). As part of his critique of past views, Strawson claims that our responses to others are "not merely devices we calculat- ingly employ for regulative purposes" (Strawson 2003, p. 93) (quoted in Shoemaker 2015, p. 7). In other words, responsibility responses are not just about shaping others' behavior, but are built into our connections with others and into who we are as moral, human, social beings (Shoemaker 2015, p. 7). Since Strawson, theorists taking his general "quality of will" approach have characterized "will" in different ways. Shoemaker thinks each characterization is deficient on its own. Thus, he claims that there are three qualities of will: quality of character, quality of judgment, and quality of regard. The kinds of responsibility associated with each are attributability (character), answerability (judgment), and accountability (regard) (viii). Let's look at each of these in turn.

First, let's discuss attributability. According to Shoemaker, "to have quality of character, an agent must have cares, commitments, or care-commitment clusters (a deep self) that are expressed in the agent's attitudes" (Shoemaker 2015, p. 59). When an agent expresses his character through his action, we can hold him responsible in the attributability sense – we can, for example, appropriately respond with disdain or admiration for the character that is expressed). How might this apply to children?

The first question is whether children have cares or commitments that they express when they act. It seems fairly clear that children have cares, even before they are able to reflect on whether these cares are ones they ought to have, and before they are able to decide what their commitments are. As Agnieszka Jaworska claims

there is ample evidence that children as young as two and three are capable of caring. Developmental psychology research shows that around age one-and-a-half children begin to “act constructively to relieve another’s distress,” both at home and in laboratory playrooms. (Jaworska 2007, p. 530)

Jaworska also discusses other kinds of cares that children have – wanting to do things on one’s own, or caring about particular objects (Jaworska 2007, pp. 530–531). She argues that such “carings” are more closely tied to the self (what she calls internality) than are our evaluations and other attitudes (Jaworska 2007, p. 537). Jaworska identifies “caring” as a complex secondary emotion.²¹ Shoemaker, who similarly characterizes caring as a grouping of emotional dispositions, agrees that our cares can be expressive of our selves (note the “cares, commitments, *or* care-commitment clusters” cited above. My emphasis). He claims that sometimes our commitments express our deep selves, but other times what expresses our deep selves is not what we have “authorized” upon self-evaluation (Shoemaker 2015, p. 55). And one’s cares may not always arise from evaluation (they may even conflict with one’s evaluative judgments. See Shoemaker 2015, p. 55. See also Jaworska 2007). So it would seem that children are not barred from attributability merely on the basis of being too young to evaluate or to have commitments.

But it is one thing to understand that children can have cares that in some way constitute who they are. It is another to think that children are responsible when they express these cares through their actions. We are most likely comfortable genuinely admiring the two or three-year-old who expresses his loving nature by acting gently with his baby brother. But are we ready to disdain the three-year-old who expresses a less kind disposition when he makes the baby cry? If not, why not?²² Young children clearly have distinct personalities, cares, preferences, and dispositions. But does a very young child have a moral *character*? It is not clear to me what we ought to think here. Even if we grant that not everything associated with our deep selves must be consciously authorized, there is nonetheless something uncomfortable about granting attributability to very young children. But it is not easy to say why. We might think character requires historical conditions, such as having developed a character. But Shoemaker wants to deny history as relevant to attributability (Shoemaker 2015, pp. 191–214). Or we might think that having a character requires more sophisticated cognitive skills. But Shoemaker suggests that dementia patients can be attributability responsible for behavior stemming from “new” character traits – based on new commitments and cares – that come about as the result of the illness (Shoemaker 2015, p. 213). That these new “care-commitment” clusters are the result of cognitive impairments is not, according to Shoemaker, sufficient for denying attributability (Shoemaker 2015, p. 213). So perhaps if we are uncomfortable holding such young children attributability responsible, we need to argue against Shoemaker in favor of historical conditions or in favor of requirements for character that young children are unable to satisfy.

Now for answerability. Answerability involves quality of judgment, which requires that “an agent’s attitudes [are] generally governable by the agent’s judgments about why they are more worthy than (some relevant) others, i.e., judgments about “instead of” reasons” (Shoemaker 2015, p. 82). Answerability connects to some of what was discussed above with regard to receptivity to reasons. For Shoemaker, receptivity involves the ability to provide “instead of” reasons. He appeals to a Scanlonian two-stage approach whereby we first decide that something is a reason and then determine the worth of this reason in relation to others (Shoemaker 2015, p. 186). This ought to remind us of the psychologists’ claim (discussed above) that children do not weight moral reasons in the way that adults do when making moral decisions. With respect to children, Shoemaker says:

we could thus explain children's burgeoning answerability by pointing to their burgeoning ability to recognize an increasing variety of "reasonish" facts. At a certain point in their development, for instance, they come to recognize that the fact that it will leave stains on the wall is actually a reason *against* drawing on it with a permanent marker. (Shoemaker 2015, pp. 76–77)

Shoemaker also characterizes answerability in terms of "applying abstract principles about wrongness to concrete judgments of moral worth" (Shoemaker 2015, p. 189). It appears that children can do this at relatively early ages, at least in general. But complex moral circumstances will be more challenging for them in terms of weighting moral reasons properly. This improves throughout development, but there does not seem to be an easy way to claim that at a particular stage a child can either do this or not. Instead, it seems more likely that it will depend on a number of issues regarding the circumstances of a particular action or choice.

Finally, let us look at accountability. Quality of regard (accountability) requires being capable of either evaluational empathy or emotional empathy (Shoemaker 2015, p. 113). Evaluational empathy involves "coming to see facts about others' (or the agent's own) normative perspectives as putative reasons in the agent's own normative deliberations" whereas emotional empathy involves feeling what another feels (Shoemaker 2015, p. 113). Shoemaker believes that both kinds of regard involve "status-recognition" – i.e., the recognition that "the target of my empathy *has* a normative perspective" (Shoemaker 2015, p. 101). Very young children react to those in distress (as discussed above). Is this emotional empathy, as Shoemaker understands it? It seems to me that the answer will depend on whether it involves the recognition of the perspective of the person who is being empathized with. It is not easy to discern when this is possible for children. But perhaps relevant is the evidence that even before age 2, children begin to be able to take on the perspective of another and get better at it in various ways as they develop.²³ So it would seem that children are capable of regard at fairly young ages. Should these young children be held accountable?

In some scenarios with young children, epistemic conditions will not be satisfied, so we will not hold the child responsible. If a child did not know that his behavior was harmful, for instance, his engaging in it does not indicate poor quality of regard. But the question is what to think about accountability in the young child whose behavior *does* demonstrate poor quality of regard. While it might seem troubling to hold a young child accountable, it may help to consider what it means to do so. In general, the appropriate response when someone shows disregard is anger, and anger is essentially communicative (Shoemaker 2015, p. 104): "what it aims to communicate is a demand for acknowledgment" (Shoemaker 2015, p. 112). So whether it is appropriate to hold someone accountable may depend on whether "there is uptake," that is, whether the person understands the request for acknowledgment and is able to respond (Shoemaker 2015, p. 112). It does not seem unreasonable to think that even preschool age children can satisfy this demand, at least in some cases. If a preschooler kicks her father, for example, it does not seem inappropriate for the father to demand that his child acknowledge why he is angry: "That hurt. How would you feel if someone kicked you? Wouldn't that make you mad?" This of course does not entail that harsh treatment is warranted by the anger, just that it is appropriate for the father to communicate (and request acknowledgment) that he has been disregarded.^{24,25}

In my view, Shoemaker's approach is illuminating because it provides resources for conceptualizing whether and in what senses an agent is part of the responsibility community. Applying it to children is helpful because it highlights different dimensions along which children might be deficient. For example, a child capable of emotional empathy might be accountable but not answerable. Perhaps he is unable to fully recognize "other-regarding

reasons” but is able to respond with emotional sensitivity to someone else’s suffering (Shoemaker 2015, p. 114). And depending on what we think about the development of character, a child may be accountable before we can attribute things to his character, or the child may be attributability responsible before being answerable.

5 Degrees of Blame

Thus far, we have considered a number of ways to assess whether children meet the requirements for moral responsibility, whether we have in mind a monistic or pluralistic conception. But is there more to the story? Aren’t there cases in which all conditions are satisfied, yet we are reluctant to respond in the same way as we would were the behavior performed by an adult? Comparative assessments seem to yield differences. If I behave in a particular way knowing that I ought not to, I seem more blameworthy than the child who performs the same intentional action. To put this in Shoemaker’s terms, we may be ready to say that a non-adult agent is part of the responsibility community because she meets the standards of attributability, answerability, and accountability. But we may yet feel that the appropriate responses are not on a par with the responses we should have for an adult. In other words, our ambivalence about children might not always fall cleanly upon his divisions.²⁶

One way to look at this is to regard responsibility as a scalar rather than as a threshold notion (Vargas unpublished). As Alfred Mele suggests, children seem to become more responsible as they develop (Mele 2006). But another possibility, subtly but importantly different from the first, is to regard *responsibility* as a threshold notion, but regard appropriate responses of *blame* (or praise) as scalar (Vargas unpublished, p. 16). So, for example, a child may be part of the moral responsibility community but less blameworthy for a particular action. This sort of account may be very useful for making sense of our ambivalence about children insofar as it allows us to acknowledge diminished blameworthiness even when a child or adolescent “knows better” and satisfies all the other basic responsibility conditions.

What might diminish a responsible child’s blameworthiness? One promising answer centers on difficulty. Dana Nelkin argues that the more difficult it is to do what one ought, the less blameworthy one is (Nelkin 2014. See also Faraci and Shoemaker 2010, 2014). My proposal is that acting as they ought is more difficult for children and adolescents than it is for us, at least in many cases. There may be occasions in which an adult has a harder time of it, but this need not do damage to the proposal under consideration. The notion of difficulty will always need to be indexed to particular circumstances. It is just that one’s developmental stage will always be *part of* one’s circumstances.

Why is it more difficult for children and adolescents? We have already touched on some areas where children have more difficulty. They may have a harder time ranking their reasons. They may have a harder time controlling themselves.²⁷ They may be at a disadvantage with respect to self-evaluation. There are most likely lots of different underlying factors that help us account for why it is harder for them. But I want to explore the development of narrative capacities as an important element.

Both philosophers and psychologists have done a lot of work on (what I call) narrative capacities, loosely understood as the abilities we have to tell and understand stories. Interestingly, according to psychologists, while children have some solid narrative skills by age 5, it is not until ages 9–11 that they “reach almost adult-level performance” in terms of structuring a narrative (Habermas and Bluck 2000, p. 752). And various elements of narrative coherence increase through adolescence. Notably, it is not until adolescence that children develop

a kind of narrative self or “life story model of identity” (see for example, Habermas and Bluck 2000; McAdams 2001; Bohn and Berntsen 2008; Fivush 2011). It is during adolescence that children start wondering and asking about their own identities. They begin thinking in more depth about the causal and thematic threads connecting their life experiences, such as how these experiences have helped to shape who they are (Habermas and Bluck 2000; McAdams 2001).²⁸

The narrative-self capacity appears to be a product of not only cognitive maturation but also social and emotional factors. It requires getting to a stage of life and social development that motivates self-directed questions about one’s identity. This is important because it reveals the mistake in thinking about children’s moral development in entirely intellectual terms. Shoemaker points out that it is a mistake to compare adults with mild intellectual disabilities (MID) to children (as has often been done). His reasoning is that adults with MID are often more mature along emotional and cognitive dimensions, in part because they have more experience and knowledge (183–184). I assume that the experience he cites includes the social and interpersonal. Although Shoemaker is not making the same claim about narrative that I am, his point seems to fit with what I suggest here. It is plausible to assume that many adults with MID have reached the maturational stage of thinking about themselves in narrative terms, whereas children have not.

Thus, from the standpoint of psychology, narrative skill develops gradually throughout childhood. Philosophically speaking, narrative is essentially about meaning. It is, as some have called it, a “meaning-affecting” relation (see Rosati 2013). It’s about what an event or action might mean to us based on its place in the story. So the development of narrative skill is the development of the ability to understand and create these meaning-affecting relations.

Deficiencies in narrative capacity plausibly have an important bearing on moral responsibility in some of the following ways. The ability to think in narrative terms may be important for controlled behavior. At the most basic level, I may need to see the connections between myself now and myself later – what I will want or need or feel. This is a rudimentary narrative capacity because it involves certain relations of connectedness (see Griffith 2019, 2020). Narrative reasoning may also be important for understanding human behavior and being able to take up the perspective of another.²⁹ These skills are also important to acting morally (note their connection to Shoemaker’s notion of regard). But it is perhaps when we look at the most robust narrative capacity – the full fledged life story capacity – that we can see an important difference between children and typical adults. It may be here that someone without this full capacity will have relatively more trouble ranking reasons because she lacks the ability to understand fully the meaning-affecting relations discussed above. In other words, she might not feel the weight of certain reasons as fully without the context of her life story as a backdrop.³⁰ Perhaps, for example, she does not fully comprehend the gravity of what she aims to do because she is only able to think of the action from within the context of relatively proximal circumstantial factors.

The narrative-self capacity is also important because it enables a child to begin to understand why she has the values that she has. Psychologist Dan McAdams (originator of the “life story model”) provides this example:

an adolescent girl may, for example, explain why she rejects her parents’ liberal political values or why she feels shy around members of the opposite sex, or how it came to be that her junior year in high school represented a turning point in her understanding of herself in terms of personal experiences from the past.... (McAdams 2001, p. 105).

A pre-adolescent child or early adolescent might have a harder time understanding her *own* values and commitments because she has not yet had the chance to reflect upon them in these causal and thematic ways and as part of her story. We can attribute these values to her, but they are in an important sense less her own until she has had the chance to evaluate them in contextual terms. The girl above who explains *why* she rejects her parents' values is beginning a process of understanding and creating her own commitments (this would be so even if this were a story of endorsement rather than rejection).³¹ I may not need to understand why I believe what I do or have the cares that I have in order to be morally responsible, but this kind of understanding could provide me with certain kinds of additional resources. It may be easier to resist peer pressure, for example, if I have thought through and understand what it is that I really value and why.

Regarding adolescence as a time of reflecting upon values and commitments is certainly not a new idea. For example, twentieth-century existentialist Simone de Beauvoir provides an insightful characterization according to which younger children believe that "human inventions, words, customs, and values are given facts, as inevitable as the sky and the trees" (Beauvoir 1948/1976, p. 35). But children eventually begin to ask questions like "what will happen if I act in another way?" and adolescents begin to understand their own role in choosing values (Beauvoir 1948/1976, p. 39). Beauvoir says: "this is doubtless the deepest reason for the crisis of adolescence; the individual must at last assume his subjectivity" (Beauvoir 1948/1976, p. 39). But today's psychology and philosophy appear to confirm what we have long suspected. My suggestion is that the narrative capacity may be able to explain why, at least in part.

6 Concluding Thoughts

In this chapter, I have tried to show that the responsibility of children is a complex issue. It is difficult to discern at what point children are full members of the responsibility community and it seems that we may have to say that their entry is not entirely linear, as they may drift in and out as they develop. But it is also important to consider that they may be appropriate candidates for praise and blame at earlier stages than we think so long as they are able to respond to reasons while also satisfying the epistemic condition; or, if we accept a pluralistic view, we might allow that they are responsible early on along some dimensions but not others. But equally important is the idea that even when children and adolescents are responsible, they are often less blameworthy because it may be more difficult for them to do what they ought to do. This may help to explain why we are ambivalent about blaming a child or adolescent even when it is difficult to discern exactly why we should be. After all, in many cases, a child or adolescent seems to know what he is doing and to have some control. Peter Ash, Chief of Child and Adolescent Psychiatry at Emory University puts it this way:

Responsibility turns not only on the present capacity to control one's actions and make sound judgments, but also on having had the opportunity to mature. As adolescents grow, they have an opportunity to reflect on their experiences, and, potentially, to break away from their environments. Consider two offenders, one an adult, and the other an adolescent, who have committed a similar crime and who have a similar sense of time perspective, intelligence, and other capacities that are taken to be elements of maturity of judgment. We should consider the adolescent as less responsible because he has not had the same amount of experience in making choices nor has he had the opportunity to reflect on his choices to the same extent as an adult. (Ash 2006, p. 148)

Ash points out the importance of the nascent evaluative abilities of the adolescent as well as the time and experience required to put them to work. Adolescents may *seem* equally capable of acting as they ought, but they are deficient in these respects, respects that can be understood partly in terms of the narrative-self capacity.

The appropriateness of our responses to children, then, must be sensitive to their status as persons and moral agents, but also sensitive to their limitations. Gaining a better understanding of these limitations also has the potential to reveal some of the subtleties of *our* responsibility. Their limitations highlight our capabilities. By seeing what they cannot do, we get a clearer picture of what we can do and why it might matter to us. As fully responsible agents, we are able to call upon a rich store of knowledge – particularly knowledge about ourselves – and on a variety of cognitive and emotional skills. Some of these skills enable us to evaluate the meaning of an event or action within its narrative context. If we find ourselves consistently more willing to fully blame those who possess these capacities than we are to fully blame those who do not, then we reveal something about what we care about when it comes to being responsible.

Notes

- 1 Some of the research that is utilized in this chapter began as a project on self-control generously funded by the John Templeton Foundation and administered by Florida State University under the direction of Alfred Mele. Many thanks to Mele and to the Templeton Foundation. Thanks also to David Shoemaker for extremely helpful comments on an earlier draft of this chapter and to Kristin Mickelson for very helpful feedback on a subsequent version.
- 2 In a similar vein, Gary Matthews says that theories of cognitive and moral development often encourage us to distance ourselves from children ... Such distancing sometimes produces a new respect for children. After all, it warns us against faulting children for shortcomings that express... immature cognitive and moral structures that are entirely normal for children of a given age range. Yet such distancing can also encourage condescension ... The condescension, though understandable, is unwarranted ... Any developmental theory that rules out, on purely theoretical grounds, even the possibility that we adults may occasionally have something to learn, morally, from a child is, for that reason, defective; it is also morally offensive (Matthews 1996, pp. 66–67).
- 3 Interestingly, Derk Pereboom cites the social and interpersonal benefits of coming to believe that we are not responsible (such as diminishing destructive anger towards one another) (see, for example, Pereboom 2002, pp. 487–488). But this at least highlights how studying responsibility has the potential to affect our relationships.
- 4 In his excellent recent book, *Responsibility from the Margins*, David Shoemaker discusses “marginal agents,” such as psychopaths and those suffering from Alzheimer’s dementia, about whom we are ambivalent when it comes to ascribing responsibility. He notes the value of looking to realistic cases of agents

who are much more akin to paradigm members of the responsibility community than fish or fowl yet who are nevertheless outside of it. Presumably, then, we could examine what precisely is different about them and so isolate the feature(s) most relevant to their outsider status in a way that reveals more determinate necessary conditions for full membership. This method should sound quite familiar, for it has been deployed by responsibility theorists stretching back to Aristotle (Shoemaker 2015, p. 6).

We might think similarly about children. One potential difference, however, is that I am not sure that children are always outside of the responsibility community the way Shoemaker has in mind.

In my view, children at certain stages of development may be fully inside the community and yet unlike us insofar as they will be more liable to fail to satisfy responsibility conditions more frequently. Or they may be unlike us insofar as they are responsible but less blameworthy. This issue and his book will be discussed more below.

- 5 I adopt this term from (Shoemaker 2015).
- 6 As David Shoemaker has pointed out in comments on a previous version of this chapter, there is an important distinction between understanding that an action is wrong and understanding why it is wrong. Although I won't discuss this distinction explicitly in what follows, I will note that it does seem to come into play in the following way. Children may have a harder time than adults weighing their reasons properly (as I discuss below). Sometimes, this might be explained by their having a diminished ability to understand why something is wrong. But I do not think the distinction between judging that an action is wrong and judging why it is wrong will, on its own do much to explain responsibility differences between children and adults. Notice that a morally responsible adult sometimes recognizes moral wrongness without being able to articulate, in any robust way, why the action is wrong and ought to be avoided: "one just doesn't do that! It's just plain wrong!" If a child makes a judgment of wrongness without the ability to articulate why, this, in itself, does not seem to be grounds for attributing diminished blame relative to the adult.
- 7 Insofar as they can correctly classify actions, we might say that psychopaths can understand which actions count as right and which count as wrong (see note 10). But even if they can classify actions, and even if psychopaths can go further than this and articulate why others think certain actions are right or wrong, there is little reason to think this means that they can actually recognize moral reasons as such. As Victoria McGeer puts it, even when psychopaths understand the motivations and beliefs of others, they do not see others as "fully real" (McGeer 2008, p. 233) and so they do not and cannot empathize. So even if they can in some sense articulate why others say it's wrong to harm, there is little reason to think they really understand why it's wrong in a moral sense. (I also discuss this in Griffith 2019).
- 8 Thanks to David Shoemaker for helpful comments on this issue. For more discussion, see Shoemaker (2017). See also Talbert (2008).
- 9 Mixed cases involve competing moral and nonmoral considerations. Moral concerns, such as considerations of fairness or of the needs of others are weighed against things like personal desires, traditions, or "conventional goals" (Pasupathi and Wainryb 2010, p. 59).
- 10 Thanks to David Shoemaker for this objection. He references the work of Walter Sinnott-Armstrong and his co-researchers. See, for example, Aharoni, Sinnott-Armstrong and Kiehl (2012).
- 11 On the other hand, some might worry that different empirical data undermines the claim that children understand moral reasons. For example, some studies show that young children regard acting immorally as impossible (see Bear and Bloom 2017. This is discussed in the "Free Will Beliefs" section below). One might then think that children do not understand the moral because they do not understand the difference between what ought not to be done and what cannot be done. This inference seems too quick to me, but more would need to be said on this issue than I can say here. In any case, this would only apply to the youngest children.
- 12 For this terminology, see Fischer and Ravizza (1998, p. 69).
- 13 One of Mischel's main claims was that success at the marshmallow test early in life correlated to success later. But recent work has called this particular claim into doubt (see Calarco 2018; Watts, Duncan and Quan 2018). Nonetheless, the point for our purposes is that these and other experiments demonstrate that children at a certain level of mental development are quite able to resist their desires.
- 14 Gopnik characterizes executive control as "our ability to suppress what we want to do now because of what we will want in the future" (Gopnik 2009, p. 148).
- 15 In comments on an earlier draft of this paper, David Shoemaker asks whether children are really deliberately trying to change the strengths of their motivations the way a typical adult agent would, rather than trying to deal with "alien" forces (much like an addict). This is a good point and might serve to demonstrate that children do not interact with their motivations the way that

- we do. Nonetheless, their behaviors seem to represent a rudimentary ability to reflect on and care about their motivations in the way a non-human animal most likely does not, even if children are not yet in a position to understand these forces as their own motivations. Furthermore, the kind of self-management they engage in is not too far removed from the kind of management adults might engage in on some occasions when they are implementing good self-control strategies.
- 16 I don't think we need to worry about the conflation of preferences and desires in the studies (children are asked about what they like rather than about what they want) since it is a reasonable assumption that their desires track these preferences.
 - 17 See also Kushnir (2012). She discusses studies suggesting that "young children do not view kindness to others as a choice" (212).
 - 18 For example, Dana Nelkin suggests that "knowing about one's own motives can be seen as one kind of knowledge needed to respond well to reasons" (Nelkin 2005, p. 201).
 - 19 See Fischer and Ravizza 1998, pp. 210–214.
 - 20 Neil Levy discusses views that dispute the need for consciousness in order to be responsive to reasons. He discusses the Huckleberry Finn example used by several philosophers, such as Nomy Arpaly (Levy 2017, p. 438). In the novel, Huck struggles with a decision of whether to help his enslaved friend Jim escape. He decides to do so but is convinced that he is acting immorally. Arpaly argues that Huck "is acting for morally significant reasons. This is so even though Huckleberry knows neither that these are the right reasons nor that that [sic] he is acting from them" (Arpaly 2002, p. 230) (I am here citing a different work than Levy cites).
 - 21 A secondary emotion differs from a primary in that the former relies on higher cognitive processing and is not just automatically triggered by "immediately sensed features of one's environment" (Jaworska 2007, p. 555). I also discuss this in (Griffith 2020).
 - 22 As Susan Wolf points out, we tend to worry more about blame, since wrongly blaming someone seems unjust, whereas wrongly praising them seems like "a harmless mistake" (Wolf 1980, p. 156). She ultimately argues for an asymmetrical thesis whereby a blameworthy agent cannot have been psychologically determined to do what he does, whereas the same need not be true of a praiseworthy agent (Wolf 1980). Thanks to Kristin Mickelson for pointing out this comparison.
 - 23 Peter Goldie discusses this, citing "false belief" studies in developmental psychology (Goldie 2007, pp. 73–75). He calls it "perspective-shifting" and uses it in an interesting argument about the importance of taking up a third-personal (external) perspective on others in order to engage in ethical thought (Goldie 2007, p. 83). While Goldie cites earlier work that puts success at such tasks at around age 3.5, more recent empirical work suggests that by changing the tasks used, much younger children (under age 2) have been successful at false-belief tasks (see Scott RM, Bailargeon R. *Trends Cogn Sci* 2017 Apr;21(4): 237–249.) Thanks to Neil Levy for making me aware of this recent empirical development.
 - 24 Shoemaker notes that responses are not the same as treatment, the latter of which would include punishment (see Shoemaker 2017, p. 664). Accepting that these young children are accountable, then, does not entail the appropriateness of punishment (one could even eschew punishment of any kind for children while accepting accountability).
 - 25 In comments on a previous draft of this chapter, Shoemaker raises the following question: Does the father really think his child has shown him disregard such that he needs acknowledgment, or is he more interested in teaching his child how to behave such that she learns not to disregard others in the future (in other words, perhaps this is better viewed as moral training than moral engagement)? It likely depends on the details of the case (e.g., whether the child lashed out in anger without thinking, or whether the child intentionally sought to cause pain). In any case, it seems conceivable that even preschoolers are capable of (at least) a rudimentary form of evaluational empathy since they appear to have all the prerequisites to some degree (empathy, knowledge of moral norms, ability to understand the perspective of another). But even if this is incorrect, it seems fairly clear that older children would be capable of disregard in various circumstances.
 - 26 This is not meant as an objection to Shoemaker, since I don't think he's suggesting that the tripartite model explains every kind of difference or ambivalence that we might want to explain. Fur-

- thermore, his view can accommodate the appropriateness of different responses based on factors beyond the requirements of each kind of responsibility.
- 27 There are a number of interesting reasons why children may have more trouble with self-control. Very young children are not yet able to fully engage in mental time travel, whereby we project ourselves into the future to think about what we will want then (see Gopnik 2009. For more on the importance of mental time travel for self-control and responsibility, see Kennett and Matthews 2009). Children may have more difficulty holding their goals before them in working memory, which could make it harder to stay on task (Hofmann et al. 2011, p. 211). And interestingly, teenagers may in fact have stronger (but not irresistible) motivations to engage in certain kinds of problematic behavior than either younger children or adults. See, for example, Somerville, Jones, and Casey (2010) and Casey and Caudle (2013). So they may have a harder time controlling themselves than younger children do under certain circumstances. I argue elsewhere (and discuss briefly below) that narrative capacities factor into self-control in significant ways (my discussion includes the claim that mental time travel is essentially narrative. See Griffith 2020). As discussed below, children are deficient with respect to narrative capacities.
 - 28 Habermas and Bluck (2000) discuss the narrative self capacity in terms of different kinds of coherence, arguing that causal and thematic coherence do not develop until adolescence or later.
 - 29 Daniel Hutto argues that we learn folk psychology through narrative practices (Hutto 2007). Peter Goldie discusses perspective in (Goldie 2007). See note 23 above.
 - 30 I develop this point in terms of reasons responsiveness in (Griffith 2019).
 - 31 Huck Finn (and Nomy Arpaly's excellent discussion) comes up again here. Recall that Huck is a young adolescent. Notice that his stated values are the "ready-made" values of his culture – an enslaved person is property and must be returned (hence his thought that he is sinning by helping him escape). Huck thinks these are the values, even though he can't bring himself to endorse them through his actions. The values that he acts upon are values we applaud – treating other persons as ends in themselves and treating them with dignity, respect, and compassion. Huck is a good example of someone on the cusp of embracing his own subjectivity, not quite yet able to articulate that the cultural values are not really his values, even though we know that they are not. But he is getting closer. By most accounts, Huck is responsible and praiseworthy for his action. For our purposes it is important to imagine how we might have reacted had his decision gone the other way. Would we have blamed him? I think that we might have, but the degree of blame would have been diminished given how hard it is for him to understand his own commitments. One interesting question that this case raises is whether difficulty introduces an asymmetry between praise and blame. Given the plausible claim that it was more difficult for Huck to do the right thing (because of his difficulty understanding his own values and commitments), might we see him as more praiseworthy? Does this mean that children and adolescents will often be more praiseworthy than we are for the same behavior? See Faraci and Shoemaker (2010, 2014) for discussion of this issue and, in particular, the Huck Finn example.

Bibliography

- Aharoni, E., Sinnott-Armstrong, W., and Kiehl, K.A. (May 2012). Can psychopathic offenders discern moral wrongs? a new look at the moral/conventional distinction. *Journal of Abnormal Psychology* 121 (2): 484–497.
- Arpaly, N. (2002). Moral worth. *The Journal of Philosophy* 99 (5): 223–245.
- Ash, P. (2006). Adolescents in adult court: does the punishment fit the criminal? *The Journal of the American Academy of Psychiatry and the Law* 34: 145–149.
- Bear, A. and Bloom, P. (2017). Born free? children's intuitions about choice. In: *Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), 477–484. New York: Routledge.
- Beauvoir, S.de. (1948/1976). *The Ethics of Ambiguity* (trans. Bernard Frechtman). New York: Citadel Press.

- Bohn, A. and Berntsen, D. (2008). Life story development in childhood: the development of life story abilities and the acquisition of cultural life scripts from late middle childhood to adolescence. *Developmental Psychology* 44 (4): 1135–1147.
- Calarco, J.M. 2018. Why rich kids are so good at the marshmallow test: Affluence – not willpower – seems to be what's behind some kids' capacity to delay gratification. *The Atlantic*. <https://www.theatlantic.com/family/archive/2018/06/marshmallow-test/561779>.
- Casey, B.J. and Caudle, K. (2013). The teenage brain: self-control. *Current Directions in Psychological Science* 22 (2): 82–87. doi: 10.1177/0963721413480170.
- Duckworth, A.L., Gendler, T.S., and Gross, J.J. (2014). Self-control in school-age children. *Educational Psychologist* 49 (3): 199–217. doi: 10.1080/00461520.2014.926225.
- Faraci, D. and Shoemaker, D. (2010). Insanity, deep selves, and moral responsibility: The case of JoJo. *Review of Philosophy and Psychology* 1(3): 319–332.
- Faraci, D. and Shoemaker, D. (2014). Huck vs. JoJo: moral ignorance and the (A)symmetry of praise and blame. In: *Oxford Studies in Experimental Philosophy* (ed. J. Knobe, T. Lombrozo and S. Nichols), vol. 1, 7–27. New York: Oxford University Press.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.
- Fivush, R. (2011). The development of autobiographical memory. *Annual Review of Psychology* 62: 559–582.
- Frankfurt, H. (January 14 1971). Freedom of the will and the concept of a person. *The Journal of Philosophy* 68 (1): 5–20.
- Goldie, P. (2007). Dramatic irony, narrative, and the external perspective. In: *Narrative and Understanding Persons* (ed. D.D. Hutto), 69–84. New York: Cambridge University Press.
- Gopnik, A. (2009). *The Philosophical Baby: What Children's Minds Tell Us about Truth, Love, and the Meaning of Life*. New York: Farrar, Strauss, and Giroux.
- Griffith, M. (2019). Narrative capacity and moral responsibility. *Social Philosophy & Policy* 36 (1): 93–113.
- Griffith, M. (2020). Children, responsibility for self-control failures, and narrative capacity. In: *Surrounding Self-Control* (ed. A.R. Mele), 142–163. New York: Oxford University Press.
- Habermas, T. and Bluck, S. (2000). Getting a life: the emergence of the life story in adolescence. *Psychological Bulletin* 126: 748–769.
- Hofmann, W., Friese, M., Schmeichel, B.J., and Baddeley, A.D. (2011). Working memory and self-regulation. In: *Handbook of Self-Regulation*, 2e (ed. R.F. Baumeister and K.D. Vohs), 204–225. New York: Guilford Press.
- Hutto, D. (2007). The narrative practice hypothesis: origins and applications of folk psychology. In: *Narrative and Understanding Persons* (ed. D.D. Hutto), 43–68. New York: Cambridge University Press.
- Jaworska, A. (2007). Caring and internality. *Philosophy and Phenomenological Research* 74 (3): 529–568.
- Kennett, J. and Matthews, S. (2009). Mental time travel, agency and responsibility. In: *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives* (ed. M. Broome and L. Bortolotti), 327–349. New York: Oxford University Press.
- Kushnir, T. (2012). Developing a concept of choice. In: *Rational Constructivism in Cognitive Development* (Series ed. J.B. Benson, Vol. ed. F. Xu and T. Kushnir), 193–218. Elsevier Inc.: Academic Press.
- Kushnir, T. et al. (2015). Developing intuitions about free will between ages four and six. *Cognition* 138 (2015): 79–101. doi: 10.1016/j.cognition.2015.01.003.
- Levy, N. (2017). Empirical perspectives on consciousness and its relationship to free will and moral responsibility. In: *Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), 434–443. New York: Routledge.
- Matthews, G. (1996). *The Philosophy of Childhood*. Cambridge, MA: Harvard University Press.
- McAdams, D.P. (2001). The psychology of life stories. *Review of General Psychology* 5 (2): 100–122.
- McGeer, V. (2008). Varieties of moral agency: lessons from autism (and psychopathy). In: *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotions, Brain Disorders, and Development* (ed. W. Sinnott-Armstrong), 227–257. Cambridge, MA: MIT Press.
- McKenna, M. (2017). Reasons-responsive theories of freedom. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), 27–40. New York: Routledge.
- Mele, A. (2006). *Free Will and Luck*. New York: Oxford University Press.

- Mele, A. (2012). *Backsliding*. New York: Oxford University Press.
- Nelkin, D.K. (2005). Freedom, responsibility and the challenge of situationism. *Midwest Studies in Philosophy* XXIX: 181–206.
- Nelkin, D.K. (2014). Difficulty and degrees of moral praiseworthiness and blameworthiness. *Noûs* 2014: 1–23. doi: 10.1111/nous.12079.
- Pasupathi, M. and Wainryb, C. (2010). Developing moral agency through narrative. *Human Development* 2010 (53): 55–80. doi: 10.1159/000288208.
- Pereboom, D. (2002). Living without free will: the case for hard incompatibilism. In: *The Oxford Handbook of Free Will* (ed. R. Kane), 477–488. New York: Oxford University Press.
- Rosati, C. (2013). The story of a life. *Social Philosophy and Policy* 30 (1–2): 21–50.
- Shoemaker, D. (2015). *Responsibility from the Margins*. New York: Oxford University Press.
- Shoemaker, D. (2017). Marginal agents and responsibility pluralism. In: *The Routledge Companion to Free Will* (ed. K. Timpe, M. Griffith and N. Levy), 656–668. New York: Routledge.
- Somerville, L.H., Jones, R.M., and Casey, B.J. (2010). A time of change: behavioral and neural correlates of adolescent sensitivity to appetitive and aversive environmental cues. *Brain and Cognition* 72: 124–133.
- Talbert, M. (2008). Blame and responsiveness to moral reasons: are psychopaths blameworthy? *Pacific Philosophical Quarterly* 89: 516–535.
- Vargas, M. (unpublished). Less than fully responsible: difficulty and degrees in shared and mitigated responsibility.
- Wolf, Susan (1980). Asymmetrical Freedom. *J Philosophy* 77 (3): 151–166.
- Watts, T.W., Duncan, G.J., and Quan, H. (2018). Revisiting the Marshmallow test: a conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science* 29 (7): 1159–1177. doi: 10.1177/0956797618761661.

The Epistemic Condition of Moral Responsibility

PHILIP ROBICHAUD

1 Introduction

Philosophical discussions about moral responsibility have traditionally been comprised of arguments about freedom and control. Specifically, philosophers have argued about what kind of freedom or control an agent must have in order to be morally responsible. Agents who episodically lack the relevant kind of control in the performance of a bad or wrong action are thought to be excused or, if their lack of control is non-episodic, exempt from moral responsibility. Until recently, less attention had been given to the epistemic condition of moral responsibility, which concerns questions about what an agent needs to know or believe in order to be morally responsible. The core idea is that, just as the lack of control can impact an agent's moral responsibility, certain instances of ignorance can do the same. Blameless and unavoidable ignorance concerning the potential harm that would result from one's actions would plausibly excuse one from moral responsibility for realizing that harm, even if there was no question regarding whether the agent was in control of what she was doing at the time. Of course, not all ignorance is exculpating, and accounts of the epistemic condition distinguish between ignorance is exculpating and ignorance that it is inculpat-ing. If, for example, I simply didn't know that the beer I offered you had been tainted at the brewery, then I am not morally responsible if you get sick. The same cannot be said for agents who act from culpable ignorance. If I knew that there was a problem of toxic beer coming from this brewery and if I knew that I could easily assess whether the bottle was tainted, but I simply failed to do so, then, intuitively, I would be blameworthy if you get sick. My ignorance in this case seems culpable because I knew your health was at stake, and I knew that my ignorance about whether you would be drinking tainted beer could have easily been avoided. It seems fair or reasonable to blame me for making you sick in the latter case but not the former. Other kinds of ignorance are less amenable to such a straightforward analysis. Consider the friend who forgets your birthday. Is their occurrent ignorance that today is your birthday exculpating? Or consider Aristotle. Was his ignorance of the irrelevance of gender to moral

equality inculcating? An account of the epistemic condition should explain the difference between blamelessly ignorant agents and culpably ignorant ones in a way that yields plausible answers these and other related questions.

In this chapter I will canvas the central positions in debates about the epistemic condition and discuss some of the core theoretical and applied questions that concern the impact ignorance has on moral responsibility. In Section 2, I discuss the structure of the epistemic condition and describe how the foreseeability of ignorance is relevant to the question of whether the epistemic condition is met. In Sections 3 and 4, I describe two of the main strands in the debate. Section 3 focuses on “volitionism,” which is a set of views unified around the claim that, roughly, culpable ignorance must trace to an instance of knowing or witting belief mismanagement. Section 4’s focus is a less unified set of “non-volitionist” views that don’t have this requirement and are, to this extent, less restrictive when it comes to the conditions that ignorant agents must meet in order to be morally responsible. Section 5 takes a different approach and explores how the topics of moral ignorance, difficulty, praiseworthiness, and luck are relevant to the epistemic condition. In the final section, I discuss some issues in what might be called the applied ethics of the epistemic condition.

2 The Structure of the Epistemic Condition; Tracing and Foreseeability

Before discussing the content of the epistemic condition, it’s worthwhile to give some attention to its structure. The formulation of the epistemic condition should, of course, be met for non-ignorant agents. More precisely, it should be met by agents whose epistemic states regarding certain normative and non-normative facts meet the relevant standard, whether it’s knowledge or justified belief. A standard formulation would be

EC-1: An agent meets the epistemic condition of moral responsibility if and they are non-ignorant of certain morally-relevant features of their action, including, for example, the actions likely consequences and its deontic status.

There is much to discuss about the elements of EC-1, some of which will come up below. If I insult someone in the full awareness that their feelings will be hurt and that it’s wrong, then EC-1 clearly holds. However, as mentioned above, ignorant agents can also meet the epistemic condition, and so a second disjunct must be added to EC that captures this.

EC-2: An agent meets the epistemic condition of moral responsibility if and only if they are (a) non-ignorant of certain morally-relevant features of their action, including, for example, the actions likely consequences and its deontic status; or (b) they are ignorant about certain morally-relevant features of their action, but the agent’s ignorance is inculcating.

Thus, on EC-2, ignorant agents can also meet the epistemic condition so long as their ignorance is inculcating. In the example above, my ignorance that the beer I offered you was toxic and would lead to your illness was intuitively inculcating given what I knew both about the chance that it might be toxic and about how to find out if this is the case. At the time of offering the beer I’m ignorant of some normatively salient feature of my action, but the fact that my ignorance traces to some objectionable and avoidable risk-taking on my part seems incompatible with the claim that this ignorance constitutes an excuse. The structure of the epistemic condition, thus, mirrors ways of establishing that an agent’s temporary loss of the sort of control that is necessary for moral responsibility can also be inculcating. At the time

of action, both agents who lack the right kind of control and agents who lack knowledge of relevant features of their actions can thus be said to meet both the control and epistemic conditions of responsibility at the time of their action, as long as their deficits of control and knowledge are inculcating rather than exculcating.

Given this structure of the epistemic condition, it's no surprise that much of the discussion regarding the epistemic condition has focused on different articulations of the conditions under which ignorance can be inculcating. Consider Smith's original discussion of culpable ignorance (1983) according to which ignorance is inculcating if and only if it is a foreseen result of some "benighting" action or omission. Ignorance that is traceable in this way to some benighting action, such as a witting failure to conduct the required investigations, would have the right kind of inculcating etiology. Zimmerman (1997, pp. 420–421) and Rosen and Levy (2009) also accept that only foreseen ignorance is inculcating. However, a weaker account would maintain that ignorance that is *foreseeable* though perhaps not foreseen can also be inculcating. One of the central reasons for relaxing the conditions is that many cases of intuitively blameworthy ignorant action do not involve the witting mismanagement of one's beliefs. On these views, ignorance that traces to negligent rather than reckless belief mismanagement can be inculcating even in the absence of clear foresight. An alternative position is that ignorant agents can be blameworthy even if it's not even foreseeable that some ignorance-producing action or omission will result in ignorance. For example, Vargas (2005, pp. 275–277) notes that someone may be unable to foresee that, say, choosing to nurture certain mildly problematic character traits will over time desensitize them to certain moral reasons and make them more likely to falsely believe that some particular, future selfish actions will be permissible. He argues that such an agent's ignorance is inculcating even though they did not and could not foresee that it would result in this ignorance about the permissibility of particular downstream actions.

These three options regarding the conditions for ignorance to be inculcating have led to a rich discussion of how to *trace* an agent's ignorance at the time of acting to some prior act of the agent's that resulted in their ignorance. Notable developments include the following. Fischer and Tognazinni (2009) defend the foreseeability view by arguing that, rather than being able to foresee particular tokens of ignorant wrong action, agents must only be able to foresee that certain ignorant act types (ex: a selfish act) would occur. In turn, Shabo (2015) argues that the foreseeability of downstream ignorant action is not a requirement under any level of description, and, notably, Khoury (2012) and King (2014) both argue that tracing actually is not necessary at all. Finally, Robichaud and Wieland (2019) argue that inculcating ignorance need only trace to benighting actions that express a lack of concern for the kinds of reasons in virtue of which the ignorant action is wrong, bad, or otherwise imprudent. The tracing debate reveals that there are a variety of positions regarding just what the etiology of ignorance must be if it is to be inculcating in a way that allows agents who act on that ignorance to meet the epistemic condition. In order to enter into it more deeply, we'll focus in the following section on an influential argument for the provocative view that the epistemic condition is so strong that very few ignorant agents actually meet it. It will then be instructive to contrast it in the following section with other ways of understanding how ignorance impacts moral responsibility.

3 Volitional Accounts

As already mentioned above, Zimmerman, Rosen and Levy independently develop the view that ignorance is inculcating only if it's foreseen. They each maintain that a necessary

condition for an agent being blameworthy for some ignorant action or omission is that they are blameworthy for being ignorant that it is wrong. Call this claim **IGNORANT ACTION**. A further necessary condition for being blameworthy for ignorance is being blameworthy for whatever past action or omission resulted in their ignorance. Call this general condition on blameworthiness for ignorance, **CULPABLE IGNORANCE**. Together we have:

IGNORANT ACTION: an agent is blameworthy for some ignorant action or omission only if they are blameworthy for being ignorant that it is wrong.

CULPABLE IGNORANCE: an agent is blameworthy for being ignorant that their action is wrong only if they are blameworthy for whatever past action or omission resulted in their ignorance.

It should be clear that both conditions are distinct as regard the locus of what the agents are blameworthy for. Notice, that **CULPABLE IGNORANCE** is a claim about S's blameworthiness for some ignorance-producing act or omission. These ignorance-producing acts or omissions are potential benighting acts, if agents are blameworthy for them. But, of course, if they are to be blameworthy for these acts or omissions, they should also be wrong or at least bad, and here it's helpful to adopt Rosen's (2004) notion of procedural epistemic obligations – obligations to, say, perform acts of inquiry or avoid unreliable sources of information. These procedural epistemic obligations are straightforward moral obligations that require agents to perform certain actions or omissions that constitute well-managing of our epistemic affairs. It follows, then, that according to **CULPABLE IGNORANCE**, an agent is blameworthy for ignorance that their action is wrong only if they are blameworthy for flouting some procedural epistemic obligation. Note, however, that blameworthiness for the latter will rest in part on whether the agent knew that it was wrong to flout their procedural epistemic obligation. There are two options: either they knew this ignorance producing act or omission was wrong or they didn't. In the former case, the situation is straightforward – the agent's ignorance that her current action is wrong is inculcating given that this ignorance traces to *wittingly* flouting a procedural epistemic obligation. Returning to the example from the introduction, if an agent knows that they morally ought to look into whether the beer they are offering someone is likely to be toxic, but they don't look into it, they are blameworthy according to **CULPABLE IGNORANCE**. But, what if the agent didn't know that they were subject to the relevant procedural epistemic obligation? If they were ignorant about *this* for whatever reason, then their flouting a procedural epistemic obligation was itself an ignorant action, the blameworthiness for which would rest on whether their ignorance regarding their procedural epistemic obligations was culpable. Applying **CULPABLE IGNORANCE** again, we see that an agent is blameworthy for ignorance regarding their procedural epistemic obligations only if they are blameworthy for whatever past act or omission resulted in *this* ignorance. It's clear, now, that a regress threatens, and it can be stopped only by the agent having wittingly flouted a procedural epistemic obligation at some point. In other words, volitionists argue compellingly that ignorant agents will only meet the epistemic condition for blameworthiness if at some time they knowingly, which is to say, akratically flouted a procedural epistemic obligation.

This view, defended by Zimmerman, Rosen, and Levy in different ways, has two powerful results. First, given that akratic belief mis-management probably tends to be rare, most ignorant action will fail to have the right kind of inculcating etiology that the initially plausible principles above require. Much ignorant action likely traces to unwitting negligence, forgetfulness, carelessness, or slips, rather than the agent knowingly flouting some procedural epistemic obligation. Second, which is a point pushed by Rosen (2004, pp. 308–310), for

any given wrong and thus potentially blameworthy action, it will be very hard to for moral appraisers to be in a position to know that the agent meets the epistemic condition. We'd have to determine whether the agent's action was an ignorant act or not, and in cases of the former, reliable moral appraisal will necessitate "locating" an instance of akratic belief mis-management somewhere in the etiology of the agent's ignorance. Rosen plausibly conjectures that this will be a tall order in most cases.

Because of these challenging and revisionist implications, many have been attracted to accounts of the epistemic condition that depart in some way from the volitionists' position. The most moderate way to so depart is to retain the basic volitionist picture, but allow that the action or omission that results in ignorance can be inculcating even if it falls short of clear-eyed, akratic belief mis-management. FitzPatrick (2008, 2017) and Robichaud (2014) take this approach. FitzPatrick argues that the CULPABLE IGNORANCE thesis can be met even in cases where the agent becomes or remains ignorant due to the avoidable exercise of some epistemic vice (ex: over-confidence or callousness), and Robichaud maintains that it can be met by ignorant agents who took themselves to have *sufficient reason* to perform the act required by their procedural epistemic obligations even if they didn't take themselves to be obligated to do so (cf. Levy (2016)). For example, even if I don't know that I have an obligation to check the beer's toxicity, I might still take myself to have sufficient reason to do so. Perhaps, I mistakenly think there are sufficient reasons not to check as well (related to the onerousness of constantly checking such things). Though they don't hew to the volitionist's entire line of argument, FitzPatrick and Robichaud agree that culpable ignorance must trace to violations of one's epistemic obligations, and that such ignorance is thereby inculcating. The next section will detail accounts of the epistemic condition that depart from the core volitionist positions more substantially.

4 Non-volitional Accounts

The class of contrasting accounts is numerous. Although they all have in common the implication that non-akratic ignorant agents can meet the epistemic condition of moral responsibility, the contours of the views vary significantly. Before articulating these differences, it will be helpful to introduce a distinction first introduced by Watson (1996) between two distinct conceptions of moral responsibility. According to the *accountability* conception of moral responsibility, agents are morally responsible whenever it makes sense to hold them to account for their actions. On this conception, to appraise someone as blameworthy is to claim that blaming them for their action is appropriate. The volitionists have this sense of moral responsibility in mind, and they maintain that it would be problematic or indeed unfair to hold ignorant agents accountable unless their action traced to *akrasia*. Several other non-volitionists also have this sense of moral responsibility in mind as they theorize about the epistemic condition. These include Sher's causalist view and a so-called quality of will views. According to the *attributability* conception of moral responsibility, agents are morally responsible whenever it makes sense to see or judge that the action reflects in an important way on the agent. To appraise someone as blameworthy for an action in this sense is just to say that their wrong action can be attributed to them – it's not some kind of free floating wrong or bad state of affairs – but this need not entail that it would be appropriate for fair to blame or to hold them to account for their wrong action. Several theorists concerned about the epistemic conditions of moral responsibility have focused on developing conditions are for attributing blameworthiness to (ignorant) agents, and as will become clear below, this different focus results in distinct ways of understanding how ignorance can excuse blameworthiness and

how it is sometimes wholly consistent with it. With this distinction on the table, I will now discuss three key non-volitionist accounts of the epistemic condition.

Sher defends an account of the epistemic condition according to which agents are blameworthy for their ignorant action only if (a) their being ignorant amounts to the falling short of a moral norm that requires similarly situated agents not to be ignorant and (b) their falling short of this norm “is caused by the interaction of some combination of his constitutive attitudes, dispositions, and traits” (2009, p. 88). An example of Sher’s is the following:

Bad Joke. Ryland is very self-absorbed. Though not malicious, she is oblivious to the impact that her behavior will have on others. Consequently, she is bewildered and a bit hurt when her rambling anecdote about a childless couple, a handicapped person, and a financial failure is not well received by an audience that includes a childless couple, a handicapped person, and a financial failure. (28)

Sher maintains that Ryland is blameworthy for telling the hurtful joke despite the fact that she falsely believed that the telling of such a joke would be morally permissible in her current context. Crucially, on Sher’s account, Ryland’s ignorance is not inculpatory because she failed to perform some procedural epistemic obligation, such as investigating potential offendees who are in the audience. Rather, on Sher’s view, Ryland meets the epistemic condition in virtue of the fact that (a) she falls short of a norm requiring agents to have true beliefs about the potentially hurtful content of her anecdote, and (b) this insensitivity is caused by her own obliviousness. The etiology of the ignorance matters since if there is no moral norm requiring the agents to have the relevant belief or if the ignorance fails to trace to the agent’s constitutive attitudes, dispositions or traits, then the ignorance may function as an excuse. However, this etiology need not include the akratic mismanagement of one’s beliefs. How often ignorance is inculpatory on Sher’s view will depend on how demanding the moral norms are that govern what agents should believe or be aware of. Sher maintains that these norms are at least expansive enough to capture many non-akratic agents, such as Ryland.

A distinct non-volitionist approach to thinking about these issues focuses on the question of how ignorance relates to or is an expression of an agent’s quality of will. According to this family of views, which is prominently represented by Arpaly (2002), moral responsibility, be it for ignorant or non-ignorant action, is grounded on the nature of the “will” from which the agent acts. Arpaly understands quality of will in terms of (un)responsiveness to moral reasons (p. 71–84). Moral reasons are in turn understood as an action’s right- or wrong-making features. Roughly, an agent acts from ill will when they perform some action for the very reasons that make the action wrong (e.g.: the sadist), and an agent acts from a deficit of good will when they fail to be sensitive to or moved by the reasons that make the action wrong, even though their reasons for acting aren’t the wrong-making reasons themselves. Thus, my failure to investigate whether the beer I’m offering you is toxic, which leads to my false belief that it’s okay to offer it, may trace to my having a deficiency of good will towards your health – though I suspect that there are reasons to believe that it’s toxic, I am insufficiently responsive to the wrong-making feature of my potentially offering you something toxic. In this case, my ignorant action does not express ill will towards you, as it would if the reason I failed to investigate was the hope that the beer would be harm-inducing. In fact, it’s consistent with acting from a deficit of good will in this case that I regret having been negligent with respect to my beliefs. Still, it suffices for my blameworthiness that I lacked sufficient regard for causing you harm, and my ignorance is inculpatory given what it reveals about my insensitivity to the relevant moral reasons. Notably, my ignorance is inculpatory even though there is no instance of akratic belief mismanagement – it’s enough that my quality of will, understood in terms of responsiveness to reasons is substandard.

Of course, in the case just discussed, my ignorance concerns some straightforward empirical fact concerning the contexts of the drink. However, Arpaly develops her account by showing how it would apply in the context of moral ignorance as well. The oft-discussed case of the ancient slaveholder involves an agent who is non-ignorant when it comes to all of the non-moral facts involved with slaveholding, but who nevertheless commits the grave wrong of slave holding because they falsely believe that slaves simply aren't owed the same kinds of moral protections that others deserve. On Arpaly's view ancient slaveholder's moral ignorance would reflect a (rather extreme) deficit of good will, if it was insensitivity to the plight faced by slaves that resulted in her beliefs regarding the permissibility of holding slaves. Other cases of moral ignorance, such as straightforward sexist or racist beliefs, can also reflect such an insensitivity to moral considerations, and on Arpaly's view, this ignorance too would be inculcating.

A third non-volitionist position falls squarely in the attributionist camp, and maintains that ignorant agents are morally responsible if the ignorance and subsequent ignorant action reflects morally objectionable judgements that the agent has made. Importantly, an ignorant agent can meet this condition even if their ignorance fails to trace to witting belief mismanagement. Indeed, as Talbert makes clear, attributionists need not even require that the agent in some way be at fault for how they came to be ignorant (2012). Smith's (2005) example of someone who forgot their friend's birthday and so omits to call her to wish her well is illustrative here. Absent some special circumstance, such as stress, that would easily explain the lapse, an agent's failure to realize that it's their friend's birthday indicates the objectionable judgement that the friend's feelings are not so important. Attributionists maintain that the agent's ignorant omission in this case is blameworthy for this reason and not because, say, she flouted some duty to set a reminder or enter it into her calendar. It's enough that the ignorance attending the action reveals the agent's judgements regarding the importance of certain values and relationships. Moreover, it's not a constraint on the fairness of blaming someone that it was reasonable to expect agents to have avoided being ignorant (cf. Levy (2009)). Talbert's more recent development of this view holds that it might be fair to blame unwitting agents even when it would *not* have been reasonable to expect them to know better about the wrongness of their practices (Talbert, 2017).

5 Clusters of Issues

The previous two sections have focused on bringing into view several general takes on how ignorance can affect an agent's moral responsibility. For this and the next section I will, instead, focus on a number of more specific themes and topics that have come up in the literature on the epistemic condition. Since much of this literature has developed independently of the context of the general theories discussed above they merit a distinct treatment here. The issues that I will discuss are moral ignorance, praiseworthiness, difficulty, and luck. Nothing I say here will be decisive regarding these issues, as my aim is only to characterize the central questions these issues pose and highlight some of the answers offered in the literature.

5.1 *Moral Ignorance*

When agents are blamelessly ignorant of some morally relevant non-moral fact, then they fail to meet the epistemic condition, and they have an excuse. However, when ignorance of non-moral fact is in some way inculcating, then agents are on the hook. A natural next

question is whether the options the same for moral ignorance, which we should understand as ignorance regarding what morality requires that is not based on ignorance of non-moral facts. For example, someone is morally ignorant regarding the moral permissibility of eating factory farmed meat, if they think it's permissible to do so, but don't have any false beliefs regarding the nature of factory farming or the suffering it entails. Must we understand and explain the impact of moral ignorance on moral responsibility in the same way that we do for agents who are ignorant of morally relevant non-moral facts? Rosen (2004) has argued that there is no fundamental difference between factual and moral ignorance when it comes to the epistemic condition. The rationale for this symmetry is the existence of duties to inquire into factual and moral issues. If factual ignorance is inculcating only when it traces to witting failures to inquire into factual matters, then why not say that moral ignorance is inculcating only when it traces to witting failures to inquire into moral matters? Relatedly, Sher (2017) notes that the same factors that strike us as relevant for determining whether it's reasonable to expect agents to avoid factual ignorance – the amount and quality of evidence or the complexity of relevant reasoning – also seem relevant to determining whether it is reasonable for agents to avoid moral ignorance. This suggests that we have at least some reason to accept the symmetry between factual and moral ignorance as regards the epistemic condition.

Still, others resist this symmetrical treatment. Harman (2011) defends the view that, unlike factual ignorance, moral ignorance never excuses. On her view, ignorant agents can be blameworthy for their ignorance and subsequent unwitting actions just by virtue of having false beliefs about morality. Blameworthy agents need not have failed to conduct any moral inquiries. Indeed, one might fail see the truth of certain moral claims even if one tried really hard and exerted a lot of effort to get it right. One consideration in favor of this view is that false beliefs about morality always entail a lack of adequate moral concern (though, for some qualifications of this claim, see Wieland (2017b)). More recently, however, Harman (2017) has raised a puzzle for the view that moral ignorance never excuses. She argues that the same principle that explains why moral ignorance is inculcating delivers the intuitively wrong result for certain cases of factual ignorance.

5.2 *Praiseworthiness*

Thus far we have focused on cases where agents act wrongly and are thus candidates for blame. When we turn to cases of right or obligatory action, the question becomes what epistemic conditions, if any, apply to praiseworthiness. Assuming that other conditions on responsibility are met, one might suppose that agents who know that they are conforming to their moral obligations meet the epistemic condition of praiseworthiness. Call this the 'simple view'. Despite this view's initial plausibility, two complications are apparent. First, it is consistent with one's knowing that one, say, ought to donate to charity, that one does so only very reluctantly or from an insidious motive, such as currying favor with others. If one is not praiseworthy because of the presence of these objectionable motivational states, then the simple view is too simple. Praiseworthiness may require more than merely knowledge that one is acting rightly. Second, even absent objectionable motivational states, an agent who correctly takes herself to be acting rightly may perform some action that does not seem to merit praise. The class of relevant cases are those in which doing anything other than what is morally obligated is very objectionable indeed (in contrast to the charity case in which even agents who fail to donate as much as they should, but who still donate some amount, arguably do not act very objectionably). For example, an agent who refrains from killing her rival in cold blood and knows that this is her moral obligation, does not seem

particularly praiseworthy. If such an agent is not praiseworthy, then there must be some way to limit the set of potentially praiseworthy actions such that if an agent performs them in the knowledge that they are right, they actually merit praise. These complications suggest that the epistemic condition of praiseworthiness is more stringent than the simple view suggests. Paulina Sliwa (2017) discusses the stringency of this condition and defends an account according to which an agent is praiseworthy only if they act rightly and conceptualize it as such. This requirement is motivated both by the recognition that there is an epistemic condition on acting intentionally, such that one acts intentionally only if one performs the act under a particular conceptualization of it, and by the idea that praiseworthiness requires intentional action.

An altogether different aspect of the epistemic condition of praiseworthiness concerns ignorance cases. An agent might conform with their obligations and do so in a way that seems intuitively to be praiseworthy, yet they might hold false beliefs about what their moral obligations are. It has been argued that Huck Finn is praiseworthy for failing to turn in Jim, who is a befriended escaped slave, even though Huck fails to believe that he acts rightly (Arpaly 2002, Sher 2009, p. 142–143). On these accounts, it doesn't matter that Jim is ignorant about what morality requires. It suffices for praiseworthiness that Huck is sensitive to the moral reasons not to turn Jim in even though he does not characterize his own actions as being at all laudable. On such an analysis, the epistemic condition for praiseworthiness can be met by ignorant agents as well. Sliwa's account, by contrast, rules out praiseworthiness in all ignorance cases.

5.3 *Difficulty*

Many have argued that the difficulty of doing what morality requires can have some effect on blameworthiness (Coates and Swenson 2013; Nelkin 2016). The main issue that has to be clarified in the context of the epistemic condition is identifying the locus of difficulty in ignorance cases. If difficulty amounts to some kind of excuse for an ignorant agent, precisely which act or series of acts has to be difficult? A plausible candidate is difficulty in performing some act of inquiry that if performed would yield true beliefs, and a compelling thought is that if it would be difficult to inquire into something, then the agent who fails make sufficient investigative efforts and acts wrongly on the resulting ignorance is less blameworthy for her conduct than an agent for whom such inquiry would have been very easy. Bradford defends a view according to which certain factors that make it difficult to avoid ignorance can result in a partial or full excuse (Bradford, 2017). These factors may be circumstantial if, for example, the evidence available to one is less than reliable, but they might also be agential, if it's something about the agent's cognitive or emotional capacities that make inquiry difficult. Bradford argues that the presence of this kind of difficulty can mitigate blame as long as the factor doesn't reflect badly on the agent (p. 191). For example, someone with attention deficit disorder might be partially excused if their disorder explains why some morally relevant inquiry was cut off short. By contrast, someone who is simply callous, uncaring, and lazy would not be excused, even if their personal traits made it more difficult for them to complete the same inquiry.

Guererro (2017) discusses a different locus of difficulty in ignorance cases, namely the difficulty in seeing that one has a duty of inquiry in the first place. The thought here is that blameworthiness may also be mitigated in cases where, given particular features of the agent or their circumstances, it would be far from easy for them to even see that they have reasons to engage in the relevant inquiry. For these agents, it's not that the inquiry itself is difficult, and,

in fact, it may be rather easy. Rather, they face a difficulty in realizing that inquiry should be undertaken in the first place. Guerrero calls this “difficulty in trying” to avoid ignorance, and discusses several factors that would make it difficult to try to inform oneself. These include the degree to which changing one’s beliefs would require one (a) to depart from the norms of society or one’s subculture, (b) to make significant sacrifices in terms of the values one accepts, and (c) to think very differently from how one currently does (211). Consider, again the case of an agent who is ignorant as to the moral permissibility of eating factory-farmed meat. It may be far more difficult for agents who come from a rural and conservative ranching community to even see that the ethics of factory farming is an issue worth exploring than it would be for someone living in a community in which animal welfare and sustainability are frequent topics of discussion. Interestingly, Guererro suggests that whether this kind of difficulty mitigates blameworthiness in ignorance cases depends on whether one adopts a volitionist view or a quality of will view on the epistemic condition in general (212).

5.4 Luck

Luck has a particular application when it comes to the epistemic condition, and interesting questions are raised for different kinds of luck, namely *constitutive* and *consequential* luck. Constitutive luck is luck concerning the character traits, dispositions or virtues and vices that one comes to possess as a result of one’s genetic endowment or upbringing. One suffers from the relevant kind of constitutive luck when one finds oneself with epistemic vices, such as arrogance and dogmatism, that lead one to be ignorant about certain morally relevant facts. This kind of luck has a historical dimension: one’s epistemic vices might well be due to coming up in a social context that tolerated and perhaps even rewarded such vices. These vices can dispose agents to neglect their duties of inquiry or to weigh available evidence in a biased or otherwise problematic way, and they can be self-sustaining if they prevent the agent from realizing that they have the vice in the first place.

The question here is whether constitutive luck is compatible with blameworthiness in ignorance cases. The argument that constitutive luck is *incompatible* with blameworthiness may be based on the claim that an agent who acts wrongly in part due to bad constitutive luck either does not have the kind of control that is required for responsibility, or does not meet some historical condition of responsibility. If an agent lacked control over coming to possess a disposition to, say, ignore the advice of epistemic authorities, then she might be said to lack control over whether or not she acts ignorantly. Whether agents with bad constitutive luck actually do lack the relevant kind of control required for moral responsibility is of course controversial (see Sher (2006) and Levy (2011)).

Recent proponents of the idea that constitutive luck is compatible with blameworthiness differ concerning the degree of compatibility. FitzPatrick (2017) argues that ignorance due to bad constitutive luck is sometimes compatible with blameworthiness for failing to inquire into certain morally weighty matters. The thought is that while constitutive luck may explain why certain agents are inclined not to investigate or otherwise reflect on the permissibility of their actions, it does not rule out this possibility completely. As long as one’s social context is such that one had sufficient opportunity to do better, it may be reasonable to expect the agent to have done better. Talbert (2017) goes further, and holds that bad constitutive luck is always irrelevant to blameworthiness. Talbert argues that an ignorant agent is blameworthy if her actions stem from a lack of due regard or contempt, and he is explicit about the fact that one can meet this condition even if one’s lack of due regard is due to constitutive luck rather than one’s prior choices. On this view, even if one’s failure to inquire into some morally

significant issue is explainable by reference to bad constitutive luck, one can still be blameworthy for one's ignorant conduct whatever one's social context happens to be.

Consequential luck refers to luck concerning the consequences of one's conduct. Here, the potentially problematic impact of luck occurs only after an agent has acted ignorantly. Smith (1983) asks us to consider two agents who perform the same culpably ignorant act-type, but whose actions have different consequences. If the consequences in one case are more severe than the consequences in the other, then would we say that the former is more blameworthy than the latter? Though it is natural to think that agents can be blameworthy for certain foreseeable and avoidable consequences of their actions, it may strike some as odd that mere consequential luck could explain a difference in degree of blameworthiness between agents who acted similarly. Smith argues that one way to avoid this implication is to deny that culpably ignorant agents are ever blameworthy for their ignorant acts given that these acts are consequences of the agent's benighting acts, which are wrongful failures to investigate, and thus subject to consequential luck. Her position is that the scope of blameworthiness in ignorance cases only includes blameworthiness for the benighting act itself and not its consequences (cf Wieland and Robichaud 2017). Other philosophers have drawn an even stronger lesson from consequential luck. Peter Graham (2017) argues that careful attention to consequential luck actually gives us reason to deny that agents are blameworthy for their actions at all. The same contingent, luck-vulnerable relation that Smith takes to hold between benighting acts and subsequent ignorant ones is said to hold between an agent forming an *intention* to act and the subsequent performing of the act. According to Graham, this results in the scope of blameworthiness including only one's intention to act and not the (ignorant) act itself. This leads Graham to conclude that here is no epistemic condition for the moral responsibility of ignorant *action*, given that all we can be morally responsible for are our intentions.

6 Applications

Having discussed the epistemic condition at a general level and in the context of several more specific and related issues, I will conclude with a short discussion of what one might call the applied ethics of the epistemic condition of moral responsibility. Several recent articles have succeeded in showing how many of the theoretical frameworks canvassed above play out in the context of pressing social and ethical problems that we currently face. Although this will necessarily capture only a portion of this work, it will hopefully give the sense that the issues discussed above have wide application.

Much of our consumption of products and food involves supply chains that involve morally questionable labor and environmental practices. If we stipulate that we have individual moral obligations to buy products that are in some sense ethically manufactured and sourced, then we face the problem of determining which products have this etiology and which don't. In this context, Wieland (2017a) has provided a helpful account of what willful ignorance amounts to with respect to the conditions under which our phones, clothes, and coffee is made. According to his analysis, an agent is willfully ignorant of the sources of their products only if the reason they don't look into it is because doing so would be inconvenient for some reason or other (111). For example, it may be unpleasant to discover both that one has a history of supporting morally questionable business practices and that cheaper, more easily purchased options are to be avoided in the future. Sarch (2016) and Wieland (2019) argue that, in such cases, willfully ignorant agents are blameworthy and that they can be

just as blameworthy for their actions as witting wrongdoers. Sarch maintains that willfully ignorant agents are less blameworthy *for their wrong action* than witting wrongdoers, but that the “balance” of blameworthiness can be made up by considering their blameworthiness for not investigating in the first place. Wieland’s simpler view is that willfully ignorant consumers can be as blameworthy for their wrong actions as those who knowingly make impermissible purchases as long as the motive from which they remain ignorant is just as bad, in the sense of reflecting moral disregard, as the motive from which the witting wrongdoer would act (1418).

Somewhat relatedly, a few recent papers have examined the question of how ignorance regarding climate change should impact an agent’s blameworthiness for failing to curb one’s emissions or to perform other actions considered morally obligatory in the context of catastrophic, anthropogenic climate change. Wundisch has argued, in line with Rosen, that whether ignorance about climate change excuses depends critically the strength of one’s procedural epistemic obligations regarding climate change (Wundisch, 2017, p. 282). In cases in which agents inhabit a social milieu in which distrust of climate science and scientists prevails, we shouldn’t expect that they would exhibit preternatural moral and epistemic sensitivity or to drop everything and give themselves over to all-consuming investigation. Rather, in such cases, we would think that even some small amount of inquiry would probably suffice for meeting appropriately adjusted procedural epistemic obligations, and that remaining ignorance regarding climate change would be blameless. In previous work (Robichaud 2017), I have applied the volitionist and quality of will approaches to this context, and I argued that, although there is some convergence between these views as it concerns cases of strategic ignorance regarding climate change in which people expressly avoid reliable sources, the verdict for other kinds of cases is mixed and requires careful attention to the cases’ particulars.

A third domain in which the epistemic condition for responsibility has come up for discussion is the context of exploitation. Philosophers have defended many accounts of exploitation that are meant to capture and explain why it’s wrong, say, for an employer to offer pitifully low wages for highly dangerous work. What makes a transaction (in this case of labor for a wage) exploitative according to most accounts is that it’s unfair in some sense, and much of the debate is centered on how to establish and understand the relevant sense of unfairness (for a helpful overview see Ferguson and Steiner (2018)). Recently, Ferguson has argued that such accounts are incomplete, and that a further epistemic condition must be met in order for a transaction to qualify as exploitative, namely the would be exploiter must either believe that their action that amounts to an unfair gain is wrong, or their ignorance that it is wrong must be culpable (Ferguson, 2021). He argues that adding this epistemic condition to an account of exploitation allow us to better capture the sense that exploiters must have *mens rea* in addition to their having engaged in a transaction that is unfair. Just how the various accounts of the epistemic condition would apply in the context of exploitation is a question that calls for further exploration.

Bibliography

- Arpaly, N. (2002). *Unprincipled Virtue: An Inquiry Into Moral Agency*. Oxford: Oxford University Press.
 Bradford, G. (2017). Hard to Know, in *Responsibility: The Epistemic Condition* (eds. P. Robichaud and J.W. Wieland), 180–198. Oxford: Oxford University Press.

- Graham, P.A. (2017). The Epistemic Condition on Moral Blameworthiness: A Theoretical Epiphenomenon, in *Responsibility: The Epistemic Condition* (eds. P. Robichaud and J.W. Wieland), 163–179. Oxford: Oxford University Press.
- Guerrero, A.A. (2017). Intellectual Difficulty and Moral Responsibility, in *Responsibility: The Epistemic Condition* (eds. P. Robichaud and J.W. Wieland), 199–218. Oxford: Oxford University Press.
- Justin, C.D. and Swenson, P. (2013). Reasons-responsiveness and degrees of responsibility. *Philosophical Studies* 165 (2): 629–645.
- Ferguson, B. and Steiner, H. (2018). Exploitation. In: *The Oxford Handbook of Distributive Justice*. (ed. S. Olsaretti), 533–555. Oxford: Oxford University Press.
- Ferguson, B. (2021). Are we all exploiters? *Philosophy and Phenomenological Research*, 103 (3): 535–546.
- Fischer, J.M. and Tognazzini, N.A. (2009). The truth about tracing. *Noûs* 43 (3): 531–556.
- FitzPatrick, W.J. (2008). Moral responsibility and normative ignorance: Answering a new skeptical challenge. *Ethics* 118 (4): 589–613.
- FitzPatrick, W.J. (2017). Unwitting wrongdoing, reasonable expectations, and blameworthiness. In: *Responsibility: The Epistemic Condition* (ed. P. Robichaud and J.W. Wieland). Oxford: Oxford University Press.
- Harman, E. (2011). Does moral ignorance exculpate? *Ratio* 24 (4): 443–468.
- Harman, E. (2017) When is Failure to Realize Something Exculpatory?, in *Responsibility: The Epistemic Condition* (eds. P. Robichaud, and J.W. Wieland). Oxford: Oxford University Press.
- Khoury, A.C. (2012). Responsibility, tracing, and consequences. *Canadian Journal of Philosophy* 42 (3–4): 187–207.
- King, M. (2014). Traction without tracing: A (partial) solution for control-based accounts of moral responsibility. *European Journal of Philosophy* 22 (3): 463–482.
- Levy, N. (2009). Culpable ignorance and moral responsibility: A reply to FitzPatrick. *Ethics* 119 (4): 729–741.
- Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.
- Levy, N. (2016). Culpable ignorance: A reply to Robichaud. *Journal of Philosophical Research* 41 (October): 263–271.
- Nelkin, D.K. (2016). Difficulty and degrees of moral praiseworthiness and blameworthiness: Difficulty and degrees of moral praiseworthiness and blameworthiness. *Noûs* 50 (2): 356–378.
- Robichaud, P. (2014). On culpable ignorance and akrasia. *Ethics* 125 (1): 137–151.
- Robichaud, P. (2017). Is ignorance of climate change culpable? *Science and Engineering Ethics* 23 (5): 1409–1430.
- Robichaud, P. and Wieland, J.W. (2019). A puzzle concerning blame transfer. *Philosophy and Phenomenological Research* 99 (1): 3–26.
- Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives* 18 (1): 295–313.
- Sarch, A. (2016). Equal culpability and the scope of the willful ignorance doctrine. *Legal Theory* 22 (3–4): 276–311.
- Shabo, S. (2015). More trouble with tracing. *Erkenntnis* 80 (5): 987–1011.
- Sher, G. (2006). *In Praise of Blame*. Oxford: Oxford University Press.
- Sher, G. (2009). *Who Knew?: Responsibility Without Awareness*, 1e. New York: Oxford University Press.
- Sher, G. (2017). Blame and Moral Ignorance, in *Responsibility: The Epistemic Condition* (eds. P. Robichaud, and J.W. Wieland), 101–116. Oxford: Oxford University Press.
- Sliwa, P. (2017). On Knowing What's Right and Being Responsible for It, in *Responsibility: the Epistemic Condition* (eds. P. Robichaud and J.W. Wieland). Oxford: Oxford University Press.
- Smith, A.M. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics* 115 (2): 236–271.
- Smith, H. (1983). Culpable ignorance. *The Philosophical Review* 92 (4): 543–571.
- Talbert, M. (2012). Moral competence, moral blame, and protest. *The Journal of Ethics* 16 (1): 89–109.
- Talbert, M. (2017). Akrasia, Awareness, and Blameworthiness. In *Responsibility: the Epistemic Condition* (eds. P. Robichaud and J.W. Wieland), 47–63. Oxford: Oxford University Press.

- Vargas, M. (2005). The trouble with tracing. *Midwest Studies In Philosophy* 29 (1): 269–291.
- Watson, G. (1996). Two faces of responsibility. *Philosophical Topics* 24 (2): 227–248.
- Wieland, J.W. (2017a). Willful ignorance. *Ethical Theory and Moral Practice* 20 (1): 105–119.
- Wieland, J.W. (2017b). What's special about moral ignorance? *Ratio* 30 (2): 149–164.
- Wieland, J.W. and Robichaud, P. (2017). Blame Transfer, in *Responsibility: The Epistemic Condition* (eds. P. Robichaud, and J.W. Wieland), 281–297, Oxford: Oxford University Press.
- Wieland, J.W. (2019). Willful ignorance and bad motives. *Erkenntnis* 84 (6): 1409–1428.
- Wündisch, J. (2017). Greenhouse gas emissions and individual excusable ignorance after 1990: A study of excusable ignorance in collective action problems. *Environmental Philosophy* 14 (2): 275–315.
- Zimmerman, M.J. (1997). Moral responsibility and ignorance. *Ethics* 107 (3): 410–426.

Forgiveness and the Emotions

LAURA W. EKSTROM

Does free will skepticism have more costs to personal life than its defenders make out? What implications does the view have concerning forgiveness, mutual respect, and vulnerability in personal relationships? Are the costs of disbelief so high that we should believe – or hope – that we do have free will? My concern in this essay is the matter of the effects of free will skepticism on some of our moral practices – including forgiveness – and on our emotional lives. I'll consider moral outrage, resentment, and indignation, as well as our emotional openness and engagement in close personal relationships.

1 Background

Free will skepticism is the view that we do not have free will – in other words, the view that we do not have the power to act freely on any occasion. The sort of free will at issue is the control in acting that is required for an agent to deserve praise for a morally good action or to deserve blame for a morally wrong action.

The leading positions in the free will debate are, from my perspective, free will skepticism and event-causal libertarianism. Event-causal libertarianism is attractive because it provides conditions that would need to hold in order for us to have a kind of power that many of us have a sense of ourselves as having – the ability sometimes to act in a way that is up to us and that is not the deterministic unfolding of the distant past and laws of nature, over which we have no control – and because it does so without making appeal to an irreducible causal relation between an agent as a substance and an event that is an action or a precursor to action. I take certain arguments for incompatibilism between free will and the causal determination of all events by the distant past, together with the laws of nature, to be successful (van Inwagen 1983, 2002; Kane 1996; Ekstrom 2000; Pereboom 2001; Clarke 2003). Further, I have argued that prominent charges leveled against event-causal libertarianism – rooted in chance, luck, randomness, accidentality, and the alleged disappearance of the agent at the crucial moment of decision – can be answered successfully (Ekstrom 2003, 2011, 2019a, 2019b).

Hence one may reasonably be a compatibilist concerning free will and causal indeterminism, where the thesis of indeterminism is the claim that it is not the case that every event is causally necessitated by prior events and laws of nature. Given incompatibilism concerning free will and determinism, and compatibilism concerning free will and indeterminism, agent-causal libertarianism is a theoretical option, but appeal to a primitive agent-as-cause faces well-known difficulties. These include the problems that appeal to an agent as unmoved mover is difficult to square with an account of the timing of action, with reasons explanation of action, with a natural understanding of what causation is, and with our best scientific understanding of the way the world works. (Kane 1996; Ekstrom 2000, 2017; Pereboom 2001, 2014; Clarke 2003. For defense of agent-causal accounts, see O'Connor 2000; Lowe 2008; Griffith 2017.) Of course, the world may not conform to the conditions of the event-causal libertarian account, either – it may turn out that, in fact, there are not probabilistic causal relations between the particular events that an event-causal libertarian account says that there would need to be, in order for us to act freely – hence the suggestion that the other leading account in the free will debate is free will skepticism.

In this paper, rather than defending the cluster of contentions in the previous paragraph, I want to take up questions concerning certain implications of adopting free will skepticism.

2 Forgiveness

In an especially interesting section of his recent book, *Free Will, Agency, and Meaning in Life*, Derk Pereboom addresses the implications of his skeptical views on free will for our emotional lives and for some activities and attitudes that our interpersonal relationships typically involve. Pereboom argues, rightly I think, that a skeptical view concerning free will does not put at risk the purposefulness of our projects or the fulfillment in life that those projects can provide (2014, pp. 193–199). We can intentionally invest our time and energy in our projects – in our academic work, for instance, or in our efforts on behalf of a charitable cause – and we can have a sense of meaning in our lives stemming from our involvement in those projects, without our investments of our time and energy being freely willed. On Susan Wolf's view of meaningfulness, our lives have meaning in them when we love what is worth loving and we engage positively with what we love (2012, pp. 8–9). It seems clear that, even if we do not have free will, we can deeply care about, be gripped by, and actively engage ourselves with worthwhile projects. Pereboom also argues that, although free will skepticism does indeed undercut the rationality of expressing certain interpersonal attitudes, such as resentment and indignation, this undercutting does not do serious damage to our personal relationships (2014, pp. 178–181). In fact, he suggests, there are analogues or alternatives to those undercut attitudes that are not threatened by the skeptical view, and these analogues would be better for personal and social relationships. Pereboom expresses overall optimism that, in virtue of its implications for our emotions and relationships, adopting free will skepticism "would be better for us all things considered" (2014, p. 176).

Against Pereboom, I want to explore the suggestion that some species of moral anger at wrongs committed against oneself and against others – resentment, indignation, and moral outrage – are important to good personal relationships and, furthermore, that they are valuable attitudes to have, in that they reflect a healthy sense of self and proper self-respect. Such attitudes in response to wrong actions are also fitting as reflections of respect for others. Free will skepticism undermines the appropriateness of these attitudes, and it threatens to undercut the practice of forgiveness that sometimes follows the experience of them. Thus there are at

least somewhat serious costs to free will skepticism, and it is not clear that the prospect of widespread adoption of the view is something toward which we should have such a rosy outlook.

Let's start with forgiveness. Consider someone who is the victim of repeated poor treatment in a marriage – it could be physical abuse or numerous betrayals in the form of affairs, lies, unfulfilled promises, and emotional neglect. Suppose that the recipient of this kind of behavior consistently forgives her spouse, staying with him. It is not uncommon that we observe this phenomenon. Now if free will skepticism is true, then her spouse does not deserve condemnation or blame – he is not morally responsible in the desert sense for his physical blows or for his humiliating tirades or for his choices to pursue affairs – and for the victim of his abuse to accept free will skepticism would mean that she should see that this is so. But then here is a natural question to raise: in any recognizable sense, can she actually *forgive* him, rather than excuse or tolerate him? She might exhort him to change, and she might request better treatment, but neither of these amounts to forgiveness. It is not clear that free will skepticism can accommodate the practice of forgiveness in a consistent theoretical framework.

In response to concerns over the implications of free will skepticism for forgiveness, Pereboom notes that a wrongdoer might express sadness that he has brought harm to someone, as well as a resolution to take steps to change, and that in cases like this the victim might “forgive” in the sense of acknowledging the wrongdoer's remorse and his intention to change, while also recognizing that the wrongdoer did not take at the time of his wrongdoing – or no longer takes – his behavior to have been morally right or justifiable. In such cases the abuser and victim might both feel regret and sadness that “things turned out as they did” during the abusive attacks or the lies or the affairs, whether through causally deterministic chains of events tracing back into the distant past or by way of some causal indeterminism that does not provide sufficient agent control for desert.

However, I do not see any such transaction that occurs between two such people, or any change in their attitudes toward each other or toward the relevant actions, as something that is recognizable as forgiveness. To think that an abuser can sensibly be forgiven for abuse for which he is not morally responsible in the sense of deserving blame seems to make a mistake concerning what forgiveness is all about.

Among the number of philosophical accounts of the nature of forgiveness, one prominent view takes forgiveness to be essentially a matter of overcoming resentment on moral grounds. Jeffrie Murphy, for instance, defends this view, arguing that the overcoming of resentment central to forgiveness must be principled, rather than being merely a matter of forgetting or of inattention (Murphy and Hampton 1988; Murphy 2003). The intentional overcoming of resentment, that is, must be done for right reasons, which may include the offender's true repentance, the offender's good intentions in acting as he did, his humiliation or the extent of his own suffering due to the wrong he committed. All of these count as right reasons because, Murphy suggests, they separate the act from the agent.

A related account of forgiveness, defended by Pamela Hieronymi, takes forgiveness to be the overcoming of resentment that crucially involves a change in judgment (Hieronymi 2001). This judgment must not include excuse for the offender, and it must not include the belief that what he did to one was not wrong. Hieronymi understands *resentment* particularly as a “protest against a past action that persists as a present threat.” The threat is present because, in the absence of atonement, repudiation or apology by the wrongdoer, the wrongdoer's action makes an ongoing claim – in the public understanding or one's own – that the way one was treated was appropriate or acceptable. Resentment denies this claim. The change in judgment partially constitutive of forgiveness, then, is this change: from the initial judgment that the

wrong action stands as a threat to the subsequent judgment that it no longer persists as a threat. This change does *not cancel out* the original judgment that the wrongdoer is morally responsible for what he did.

A third account is a debt-release model, defended by, for instance, Dana Nelkin, on which forgiveness is releasing an offender from a debt owed to the victim on the basis of the offense (Nelkin 2013). On this account, when we forgive someone, we relieve him from something, an obligation he has, due to his wrongdoing, or a sort of debt he owes to us on the basis of the act he committed, without excuse, against us.

Some theorists have suggested that, in granting forgiveness, a victim of wrong re-orienta a relationship that has been disrupted or compromised by wrongdoing (Pereboom 2014, p. 189). The notion of relationship restoration, though, should not be central to the analysis of forgiveness. This is because we can forgive people, such as an abusive parent, with whom we judge it to be best to no longer have a relationship. Our reasons for ending the relationship might include protection of our own safety and psychological health, and they might include protection of the well-being of our children. Deliberately remaining out of relationship with someone does not mean that we have not forgiven that person for past wrongs. Furthermore, we can forgive people who have harmed us, such as thief or a professionally negligent physician, with whom we have had no relationship and with whom we have no interest in forming a relationship. Relationship restoration might be involved in some instances in which one individual forgives another, but they are not equivalent phenomena.

The prominent accounts of forgiveness in terms of overcoming resentment or in terms of debt release uphold in common a crucial point: forgiveness itself is not excuse. When we forgive, we maintain belief in the perpetrator's culpability for his wrong act, yet we no longer hold it against him. If I forgive you after you have harmed me, I continue to think that what you did was wrong and that you are morally responsible for it – for if your act was not wrong, or you were not responsible for it, then there is nothing to forgive, after all. To understand forgiveness as a coherent practice, we have to understand how we can coherently both cease to hold a wrong against the perpetrator and be fully aware of the perpetrator's blameworthiness in the desert sense for that wrong action, at the same time. If no one is morally responsible for anything he does, then no one incurs any genuine debts to others by way of actions for which he or she is morally responsible, and there is no fitting moral anger, indignation, or resentment to be overcome, and so forgiveness is undermined.

Pereboom explores the idea that there may be “forms of forgiveness” that are not undermined by free will skepticism and in this connection writes that “Dana Nelkin can be recruited as an ally” (Pereboom 2014, p. 188) since her account of forgiveness does not require a fore-swearing of resentment on the part of the forgiver. But as I read Nelkin's account, she is not really an ally, in that her account does require that the forgiver retain the judgment that the wrongdoer is morally responsible what he has done. She writes: “if we are to distinguish forgiveness from excuse, we must preserve the idea that the forgiver continues in her attribution of responsibility to the offender for the offense, and at the same time explain the change that constitutes forgiveness in a different way” (Nelkin 2013). Pereboom also points to Hieronymi's conception of forgiveness as withdrawal of a protest to a threat; but, while it is true that a free will skeptic could acknowledge a wrongdoer's change of heart, Hieronymi's account is not one fully friendly to the free will skeptic, since it is “uncompromising,” which means in part that, on the account, the withdrawal of the protest does not involve a backing away from one's attribution of moral responsibility to the wrongdoer.

In short, attributing moral responsibility to wrongdoers is a fundamental part of forgiveness and as such it is a fundamental aspect of human life. If no one is morally responsible in

the first place for what he has done, then forgiveness has no place. Free will skepticism hence undermines a fairly significant aspect of our interactions with each other.

Notice that the undermining I have highlighted is especially disturbing for anyone whose metaphysical outlook is theistic and whose theological system relies centrally on the notion of forgiveness. This is so whether the type of forgiveness central to one's theological outlook is divine forgiveness for human sin, or a moral mandate for us to forgive others as we have been forgiven, or both. Free will skepticism, in virtue of its implications concerning true forgiveness, sits particularly uncomfortably with such beliefs and moral directives. Free will skepticism reduces divine forgiveness for human sin to something like toleration or excuse, and it reduces a moral mandate for us to forgive others to a directive that we have a kind of sadness or regret that wrong actions happened and that we let go of bad feelings toward others on the basis of the wrongs they commit, simply "moving on" without an attribution of moral responsibility in the desert sense. (Free will skepticism sits uncomfortably with theism in other ways, too; for instance, many theists rely heavily on an appeal to human free will in attempts to rationally reconcile the existence of God with all the suffering in our world. For discussion see Ekstrom 2021.)

3 Moral Anger, Respect, and Self-respect

A view on which forgiveness is never called for seems to sit in tension with our rather widespread conviction that a tendency to be overly forgiving indicates a lack of self-respect. A victim who stands up for herself by refusing to forgive some instances of wrongdoing, conversely, demonstrates self-respect, and part of what she declares in respecting herself is that she deserves better treatment and that those who treat her poorly do not deserve her forgiveness.

Suppose we understand self-respect as an appreciation a person has for her own inherent worth in virtue of being a person, along with a desire to preserve this appreciation and a disposition to act in ways that protect or uphold it. Self-respect, on this conception, is a cluster of attitudes and dispositions concerning who we are, rather than what we do, and involves accepting one's inherent value without regard to failures and successes. A self-respecting person has proper regard for, and acts in ways that uphold and protect, her own dignity as a person. We might use the related term "self-esteem" to refer to positive feelings and thoughts a person has toward herself on the basis of her abilities or features, such as that she's good at remaining optimistic or that she has a fit physique. Self-esteem can be harmed by our perceived failures and by the mistreatment, harsh words, and rejection of others; and conversely it can be enhanced by mastering challenges and by the praise, love, and acceptance of those about whom we care. Self-esteem hence can fluctuate through time in response to such factors and as we gain or lose traits. By contrast, self-respect is something we ought to have consistently as persons.

A person who has appreciation for her own inherent worth as a person demands treatment consistent with that worth and, so far as it is in her power, does not accept behavior that treats her as less than a person of equal worth to other persons. Behaviors that treat another as less than a person of equal worth to other persons include, obviously, genocidal slaughter, the selling or purchasing of a human being to be used as a slave, and recurring physical and emotional abuse, as well as, arguably, sexual assault, and systematic betrayal and deception in the context of a committed intimate relationship the terms of which include honesty and fidelity. A hard line would be that such actions or patterns of action are unforgivable or ought

not to be forgiven by any self-respecting person. On the other hand, one might point out that demanding good treatment together with not accepting poor treatment is one thing, but forgiving poor treatment is another. Perhaps a self-respecting person can extend forgiveness to the perpetrators of grievous wrong actions or patterns of action. It is part of our common moral outlook, nonetheless, to view individuals who forgive too readily, too broadly, and too indiscriminately as, in ordinary terminology, “doormats,” as lacking in self-respect. Some cases of wrongdoing, whether in virtue of their severity or in virtue of something about the offender’s attitude or character, we view as cases in which we ought not to forgive.

A free will skeptic such as Pereboom might aim to make sense of cases of undeserved forgiveness by pointing to a lack of genuine remorse, apology, and efforts to change on the part of the wrongdoer, and he might point out that remorse, apology, and efforts to change are not undermined by free will skepticism. Of course if no one has free will, then when someone is not the least bit apologetic or repentant concerning the harm he has done to someone else, he does not deserve blame for his failure to repent, and, conversely, when someone is profusely apologetic and repentant, he does not deserve praise for this, either. The free will skeptic might contend that distinguishing cases of deserved and undeserved forgiveness does not require that anyone is morally responsible in a basic desert sense. But if there cannot be forgiveness if there is no moral responsibility in the desert sense, then there can be no cases of deserved forgiveness and there can be no cases of undeserved forgiveness if there is no moral responsibility in the desert sense. Free will skepticism then seems to undermine a distinction that is important in our moral practice, particularly in our aim to respect ourselves and in our exhortations to others to respect themselves (although there might be an analog that tracks a parallel distinction, which free will skepticism could preserve).

Notice that if we are free will skeptics, then we endeavor neither to experience nor to express certain emotional attitudes. These include moral outrage, which we might characterize as a feeling of fury together with the judgment that the object of our fury deserves this reaction of ours on the basis of an action he performed in the knowledge that it was wrong, unjust, or offensive. These attitudes also include resentment, which we might characterize as a kind of simmering negative feeling of anger toward a person together with the judgment that that person wronged one, oneself, by way of an injustice or offense on the basis of which the wrongdoer deserves one’s blame. The attitudes also include indignation, which one might characterize as a negative feeling of anger toward a person together with the judgment that that person freely wronged a person other than oneself and so deserves blame for that wrong action committed against another. Pereboom writes, “On the skeptical view, an expression of resentment or indignation will involve doxastic irrationality when it is accompanied by the belief – as in my view it always is – that its target deserves in the basic sense to be its recipient” (2014, p. 179). If we are free will skeptics, then we are committed to repudiating such attitudes, ridding ourselves of these attitudes when we do experience them, so far as we can, and limiting our expression of them. (For discussion see Shabo 2012, pp. 106–107; Pereboom 2014, pp. 179–182.)

But suppose we were to manage to do this. Would we, then, let ourselves be vulnerable and emotionally at risk in the face of someone else, in ways we typically associate with being deeply emotionally connected with another human being? Taking other persons as the proper objects of our resentment, indignation, and moral outrage, as well as our forgiveness, seems to me to be a sign of respect for them. We respect others when we treat them in ways that are consistent with their personhood. One way in which we do this is by engaging with them fully as persons, which includes making ourselves vulnerable to them by according them the power to affect us in emotionally significant ways and includes seeing them as directors of their lives rather than as tools for us to use to achieve our own ends. When we consistently

hold ourselves emotionally distant from others and treat them as things in our environment to be manipulated by incentives and penalties, we demean them.

The free will skeptic might object that this last comment is unfair. Pereboom, for instance, points out that our repudiating attitudes such as resentment and indignation toward others does not imply that we view persons merely as objects, like falling chunks of hail that might hit us on the head or as machines that need to be fixed when they malfunction. In the absence of the attitudes undermined by free will skepticism, we can still recognize persons as capable of rational reflection and capable of change in response to reasons. In this sense, we respect them and acknowledge their dignity. He points to the case of parents who react to the actions of their teenaged children not with resentment or indignation but rather with disappointment or sadness and who then endeavor to encourage their teens to change their behavior for the better. In so doing parents do not demean their teenagers or fail to care about them in a personal way.

But insofar as we think that resentment and indignation are inappropriate reactions for us as parents, it seems to me that this is true because of our role as parents, the *superiors* of those in our charge and care. The point seems not to hold in cases of equals: adult friends and lovers. As parents it is our job to withhold parts of ourselves from our children, to keep ourselves apart from them, since we are not equals. But it is our job or a vital part of our role as friends and lovers to extend ourselves – our full selves in the cases of very close relationships – as part of our respect for them, giving them the power to make us morally outraged, resentful and indignant, as well as the power to make us feel overjoyed and deeply grateful.

Again, on free will skepticism, we should seek to eliminate our attitudes of moral outrage, resentment, and indignation so far as we can, and when we experience them, we should contain them, in the sense of focusing on their moral inappropriateness and aiming to minimize their impact on our behavior. Instead of becoming or remaining angry when others hurt us or hurt other people, free will skepticism indicates that we should be disappointed or sad. However, repeatedly becoming *sad* that things have turned out the way they have – when we are harmed and when we harm others and when other people harm other people – makes us, well, rather often very sad! Sadness – pervasive and unrelenting – is depressing. By contrast, many of us find moral anger and outrage energizing; these attitudes motivate us to take positive action, to stand up for ourselves and for others. Moral anger is arguably healthy (Tavris 1989, p. 285; Haidt 2003, p. 856; Nichols 2007, pp. 417–419). It is a signal that one does not tolerate mistreatment.

Pereboom may reply that the relevant empirical evidence is mixed, and he suggests that adopting free will skepticism can make us more magnanimous, less retributive, and can lead to our having more emotional equanimity. He maintains that these are all good things. It seems to me not unreasonable to think, however, that being rather Buddha-like (or what the common caricature of being Buddha-like amounts to) is instead a kind of flattening of the emotional life, a hollowing out of it, and leads to a detachment from others that is isolating and unhealthy. It may diminish us or be detrimental to our selves. It may lessen the amount or extent of ourselves that we “put on the line” when involved in close relationships.

4 Conclusion

Some might read my remarks above as making a kind of case for compatibilism, in the spirit of Peter Strawson (Strawson 1962; Nichols 2007). Strawson defends our reactive attitudes – gratitude, love, resentment, indignation, pride, shame, and similar attitudes that we experience in response to our perceptions of others’ attitudes toward us and toward others as expressed in

their actions (and, in the case of self-reactive attitudes, in response to our own attitude toward ourselves) – as justified regardless of the truth or falsity of causal determinism. These attitudes are natural to us, Strawson emphasizes, and they are very difficult or practically impossible to shed. They are the part of the fabric of human relationships, and they give sense to our practices of praising, blaming, punishment, and reward as expressions of the reactive attitudes, without our being sensitive, Strawson argues, to metaphysical concerns over causal determination. Instead we are sensitive to concerns over the efficacy of expressing such attitudes in varying cases, such as considerations of insanity or extreme youth.

I do not take the concerns I have expressed here over some implications of free will skepticism, however, as constituting a case for compatibilism. I have here taken the strength of arguments for incompatibilism concerning free will and determinism as a backgrounded given. Others, then, might read the sections above as leaning toward a case for *revisionism* about free will. One might think that, while libertarianism is indeed supported by the intuitions of many people and by some powerful philosophical arguments, nonetheless in light of scientific evidence, or in light of the conceptual failures of libertarian accounts of free will, we should adjust our view of the sort of agent control over action needed for deserved attributions of praise and blame. We should adjust our concept of free will so that it is compatible with determinism. Manuel Vargas defends this line, although his justification of our responsibility practices is at bottom pragmatic in nature: our aim is to build better beings, to shape persons so that they act in appropriate ways (Vargas 2013). The route I have developed here instead takes the arguments for incompatibilism concerning determinism and the sort of free will required for deserved praise and blame, as well as an event-causal libertarian account of free will, more seriously, and it recommends that, should those arguments and that account be overturned, the right course is to give up on the idea that we have free will and so to give up on the idea that we are morally responsible in the basic desert sense.

As I indicated at the outset, the leading contenders in the free will debate from my perspective are event-causal libertarianism and free will skepticism. In this essay, I have aimed to motivate a leaning toward libertarianism in virtue of the costs of being a free will skeptic, particularly the implications for our emotional lives. The issue on which I have reflected is whether or not we would have to radically change ourselves on an emotional level, and whether or not we would have to cease doing some things – such as forgiving – if we were to become free will skeptics, and if we would, whether this is a good thing or a bad thing. I am not alone in thinking that acceptance of free will skepticism would have some rather profound impacts on our emotional lives and personal relationships. It seems that if we were to accept free will skepticism, then forgiving would be something it is irrational to do or to feel. If free will skepticism implies that we need to take some attitude or action that is not forgiveness but that Pereboom may describe as related to it – *holding nothing against anyone* – then the view may be objectionable on moral grounds, in that to take that attitude or course of action may undermine self-respect and respect for others. I have not said enough here to defend these worries, but perhaps at least they invite free will skeptics such as Pereboom to say more on the matter.

If these concerns over moral anger, resentment, indignation, and forgiveness are right, about even some of these important features of our personal lives, then this might form the beginnings of a pragmatic argument against accepting free will skepticism. Pereboom argues against a purely pragmatic justification for believing in free will (2014, pp. 177–178). I have not tried here to give a full pragmatic argument for accepting that we do in fact have free will, but rather only some considerations in favor of looking again at the arguments against competing positions in the free will literature, such as event-causal libertarianism. Perhaps the free will skeptic should reconsider.

Bibliography

- Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.
- Ekstrom, L. (2000). *Free Will: A Philosophical Study*. Boulder, CO: Westview Press.
- Ekstrom, L. (2003). Free will, chance, and mystery. *Philosophical Studies* 113: 153–180.
- Ekstrom, L. (2011). Free will is not a mystery. In: *The Oxford Handbook of Free Will*, 2e (ed. R. Kane), 366–380. Oxford: Oxford University Press.
- Ekstrom, L. (2017). Event-causal libertarianism. In: *The Routledge Companion to Free Will* (ed. N. Levy, K. Timpe and M. Griffith), 62–71. New York: Routledge.
- Ekstrom, L. (2019a). Toward a plausible event-causal indeterminist account of free will. *Synthese* 196 (1): 127–144.
- Ekstrom, L. (2019b). Luck and libertarianism. In: *The Routledge Handbook of the Philosophy and Psychology of Luck* (ed. Robert Hartmann and Ian Church), 239–247. New York: Routledge.
- Ekstrom, L. (2021). *God, Suffering, and the Value of Free Will*. New York: Oxford University Press.
- Griffith, M. (2017). Agent-causal libertarianism. In: *The Routledge Companion to Free Will* (ed. N. Levy, K. Timpe and M. Griffith), 72–85. London: Routledge.
- Haidt, J. (2003). The moral emotions. In: *Handbook of Affective Sciences* (ed. R.J. Davidson, K.R. Scherer and H.H. Goldsmith), 852–870. Oxford: Oxford University Press.
- Hieronymi, P. (2001). Articulating an uncompromising forgiveness. *Philosophy and Phenomenological Research* 62: 529–555.
- Kane, R. (1996). *The Significance of Free Will*. Oxford: Oxford University Press.
- Lowe, E.J. (2008). *Personal Agency: The Metaphysics of Mind and Action*. Oxford: Oxford University Press.
- Murphy, J. (2003). *Getting Even: Forgiveness and Its Limits*. Oxford: Oxford University Press.
- Murphy, J. and Hampton, J. (1988). *Forgiveness and Mercy*. Cambridge: Cambridge University Press.
- Nelkin, D. (2013). Freedom and forgiveness. In: *Free Will and Moral Responsibility* (ed. I. Haji and J. Caoette), 165–188. Newcastle upon Tyne: Cambridge Scholars Press.
- Nichols, S. (2007). After incompatibilism: a naturalistic defense of the reactive attitudes. *Philosophical Perspectives* 21: 405–428.
- O'Connor, T. (2000). *Persons and Causes*. New York: Oxford University Press.
- Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
- Shabo, S. (2012). Where Love and Resentment Meet, *Philosophical Review* 121: 95–124.
- Strawson, P. 1962. Freedom and resentment. *Proceedings of the British Academy* 48: 1–25.
- Tavris, C. (1989). *Anger: The Misunderstood Emotion*. New York: Simon and Schuster.
- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.
- van Inwagen, P. (2002). Free will remains a mystery. In: *Oxford Handbook of Free Will*, 1e (ed. R. Kane), 158–179. Oxford: Oxford University Press.
- Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.
- Wolf, S. (2012). *Meaning in Life and Why It Matters*. Princeton, NJ: Princeton University Press.

Free Will and Moral Luck

ROBERT J. HARTMAN

OHIO NORTHERN UNIVERSITY

1 Introduction

Philosophers often write about problems of free will and moral luck in isolation from one another.¹ But these problems are both centrally about control. One problem of free will concerns the difficult task of specifying the kind of control over our actions that is necessary and sufficient to act freely (cf. van Inwagen 2008). One problem of moral luck refers to the puzzling task of explaining whether and the extent to which people can be morally responsible for actions that are permeated by luck – that is, by factors beyond their control (cf. Hartman 2019a, pp. 227–229).

This chapter explicates and assesses skeptical, compatibilist, and libertarian approaches to moral luck. In the first section, I explicate prominent problems of free will and moral luck, and I highlight that what makes them distinct is largely their emphasis on different kinds of luck; along the way, I also define free will, luck, moral luck, and Nagel's categories of luck. In the second section, I summarize two leading skeptical arguments from luck, and I sketch a unified reply to them. In the third, fourth, and fifth sections, I consider and assess support and implication relations between prominent kinds of compatibilism and libertarianism, on the one hand, and causal, constitutive, circumstantial, and resultant moral luck, on the other.

2 Problems, Distinctions, and Definitions

There is no standard use of the term “free will.” But as Timothy O'Connor and Christopher Franklin (2018) note, many philosophers characterize free will as a kind of control or up-to-us-ness over choices or actions. These philosophers either identify that kind of control as the same kind of control relevant to being morally responsible for a choice (for example, Smilansky 2000, p. 16; Levy 2011, p. 1; Pereboom 2014, pp. 1–2), or, more strongly, they

A Companion to Free, Will, First Edition. Edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White.
© 2023 John Wiley & Sons, Inc. Published 2023 by John Wiley & Sons, Inc.

define freedom-relevant control in terms of the control relevant to moral responsibility (for example, Wolf 1990, pp. 3–4; Mele 2006, p. 17).² This prominent linkage of free will, moral responsibility, and control is plausible in my view, and it has the additional advantage of bringing the free will debate into direct relation to the moral luck debate, which is essentially about control and moral responsibility.

The historically dominant *problem of free will* is about whether exercising freedom-relevant control over an action is compatible or incompatible³ with its being causally determined by factors beyond the person's control. Thomas Nagel (1979, p. 35) categorizes this kind of causal determinism as a kind of luck that we may call "causal luck."⁴ So, *causal luck* occurs when an action is causally determined by factors outside of the actor's control.

Nagel, however, primarily uses other kinds of luck to formulate the *problem of moral luck*: a person is morally responsible only for what is within her control (Nagel 1979, pp. 25, 28), but, once we factor out from moral responsibility all the resultant, circumstantial, and constitutive luck – that is, factors beyond a person's control – there is nothing left to evaluate; no one is morally responsible for anything (Nagel 1979, p. 35).⁵ In what follows, I define and illustrate these categories of luck and explicate Nagel's basic skeptical argument.

Resultant luck occurs when an agent performs an action or omission with a consequence that is at least partially beyond her control (Nagel 1979, p. 28). For example, Killer is at a party and drives home drunk. At a certain point in her journey, she swerves, hits the curb, and kills a pedestrian who was on the curb. Merely Reckless is exactly like Killer, but, when she swerves and hits a curb, she kills no one. No pedestrian was on the curb. Plausibly, Merely Reckless is morally responsible only for driving recklessly. But if luck must be factored out of moral responsibility and the salient difference between Killer and Merely Reckless is a matter of luck, then Killer also is morally responsible only for driving recklessly – and so Killer cannot be morally responsible for killing a pedestrian.

Circumstantial luck occurs when it is outside of the agent's control that she faces a morally significant challenge or opportunity (Nagel 1979, p. 28). For example, No Start gets drunk and gets into her car in the same way as our first two characters, but the difference is that her car does not start. As a result, she is forced to call a cab. If No Start's engine had turned over, she would have freely driven drunk just as they did. Plausibly, No Start is morally responsible only for forming the intention to drive in a reckless way. But since the salient difference between Killer, Merely Reckless, and No Start is a matter of luck, Killer and Merely Reckless also are morally responsible only for intending to drive recklessly – and so they cannot be morally responsible for driving recklessly or killing a pedestrian.

Constitutive luck occurs when an agent's dispositions or capacities are non-voluntarily acquired (Nagel 1979, p. 28). For example, Scarred Childhood is very much like the others except that her father was killed by a drunk driver, and she subconsciously developed the policy never to drive drunk. If that traumatic experience had not occurred, Scarred Childhood would have had some different character traits and would have formed the intention to drive drunk as the others do. Of course, it is outside Scarred Childhood's control that she had the traumatic experience, and it is outside the other's control that they never experienced that kind of trauma. Plausibly, Scarred Childhood is not morally responsible for anything relevant to drunk driving. But the salient difference between the four is a matter of luck. It follows that Killer, Merely Reckless, and No Start cannot be morally responsible for anything related to drunk driving – and so they cannot be morally responsible for intending to drive recklessly, driving recklessly, or killing a pedestrian.

This skeptical argument generalizes far beyond these examples due to the way in which human beings are ubiquitously subject to resultant, circumstantial, and constitutive luck. The generalized conclusion is that no one is morally responsible for anything.⁶

These problems of free will and moral luck focus on different kinds of luck and take as premises different claims about the causal structure of human action (Hartman 2017, p. 5). The problem of moral luck, as exemplified in Nagel's skeptical argument, focuses on resultant, circumstantial, and constitutive luck, and it relies on no premise about causal indeterminism or determinism of human action. (Causal luck is typically set aside in formulating the problem of moral luck to avoid redundancy, because causal determination plays such a central role in the problem of free will.) Consider, in contrast, two of the most historically prominent skeptical arguments relating to free will. The argument for *hard incompatibilism*: (i) there must be indeterminism at the moment of choice to act freely but (ii) whatever choice occurs via the indeterministic process is lucky in a way that precludes freedom. The argument for *hard determinism*: (i) all human actions are causally lucky and (ii) an action's being causally determined by factors beyond the actor's control rules out its being a free action. These two arguments focus on causal and indeterministic luck, and each has a premise about the causal structure of human action.

This characterization of the difference between these problems is correct at least as a matter of emphasis. It is, however, worth noting that philosophers writing about free will have long been alive to problems raised by constitutive luck (see Strawson 1986, pp. 21–51; Watson 1987; Mele 1995, pp. 144–255; Smilansky 2000; McKenna 2004; Russell 2017a, 2017b). Kristin Mickelson (2019) even argues that the problems of free will and moral luck are merely different heuristic approaches to the same general problem that constitutive luck makes it impossible to act freely. In the next section, I explicate and assess skeptical arguments from constitutive luck.

Two more definitions will be helpful in the discussion to follow.

An event is subject to *luck* for some person insofar as it is influenced by factors beyond the person's control. This is the "standard" definition of luck in the moral luck literature (Hartman 2017, p. 23; for example, see Anderson 2019; Mickelson 2019; Statman 2019; Talbert 2019).⁷ But the lack of control conception of luck fails to capture at least some of our intuitions about which events are lucky and not lucky. For example, it is outside of my control that the sun rose today, and so the lack of control account of luck implies that it is lucky for me that it rose. Intuitively, however, it is not lucky for me that the sun rose (Latus 2003, p. 476). Such considerations have led some philosophers to revise the standard view of "luck" in "moral luck" (for example, see Latus 2003; Levy 2011, 2019).⁸ Nevertheless, in Hartman (2017, pp. 23–31), I argue, perhaps surprisingly, that the moral luck debate is not about luck per se but is rather about lack of control (cf. Church and Hartman 2019, p. 6). The skeptical problem to which Nagel points us can be fully explicated without using the word "luck": in brief, ubiquitous factors outside of our control so greatly influence our constitution, character, whims, deliberations, intentions, tryings, actions, omissions, and consequences that what we do and cause are not suitably up-to-us for us to be morally responsible for them.⁹

Moral luck occurs when factors beyond an agent's control partially determine her degree of positive praiseworthiness or blameworthiness (Hartman 2017, p. 2). The interesting question about moral luck is not about whether a person is praised or blamed for something affected by luck, which is how Nagel (1979, p. 26) defines it, but rather about whether she deserves to be praised or blamed for something that is subject to luck. The word "positive" rules out the idea that moral luck is responsibility-negating luck; there is no moral luck where there is no moral responsibility. The controversy, then, is about whether instances of causal, constitutive, circumstantial, and result *luck* can be instances of *moral luck*.

3 Responsibility Skepticism and Luck

Neil Levy's (2011) Luck Pincer and Galen Strawson's (1986, 1994) Basic Argument are the two skeptical arguments from luck that have garnered the most attention; both arguments centrally appeal to constitutive luck as a problem for free will. I summarize them in terminology set out in the last section and sketch a unified reply.

Levy's (2011, pp. 84–109) Luck Pincer targets history-sensitive compatibilist accounts of free will. *Compatibilist* views of free will describe an action's being causally determined by factors beyond the person's control as compatible with its being a free action. *History-sensitive* compatibilists regard a certain kind of history or the absence of a certain kind of history as a necessary condition for directly free action (Fischer and Ravizza 1998, pp. 170–239; Haji and Cuyppers 2007; Mele 2019; cf. Levy 2011). On Levy's (2011, p. 88) version, a person acts directly freely only if she has performed past actions that have strengthened, maintained, or weakened various parts of her mental dispositions in a normal way; in other words, a person must overcome her constitutive luck through a self-making process before she can act freely. The major motivation for all such history-sensitive conditions is that they protect accounts of free will from counterintuitive implications in cases of manipulation. Suppose a bad neuroscientist transforms gentle Vic's virtues to vices while he sleeps, and, when Vic wakes up that morning, he chooses to murder in a way that is wholly motivated and explained by his new vices. Is Vic blameworthy for murder? History-sensitive compatibilists say "no" – Vic is not blameworthy for murder – because he has not had a chance to strengthen, maintain, or weaken his new values.¹⁰ Vic's being vicious is just bad constitutive luck, and it rules out his acting freely.

The core of the Luck Pincer is that a person's constitutive luck cannot be undone by a process that is itself permeated with luck; the actions whereby a person takes responsibility by strengthening, maintaining, or weakening her character are themselves lucky (Levy 2011, pp. 89–97).¹¹ In particular, such actions are either constitutively or circumstantially lucky depending on the relationship between the agent's mental dispositions and her reasons for action. Suppose that a person's reasons decisively support one action over others due to the way in which her constitutively lucky mental dispositions favor that action over others. In that case, the action is constitutively lucky. Suppose instead that her reason-giving dispositions do not decisively support one action over another. In that case, lucky features of her circumstance such as mood, attention direction, or environmental influences at or near the time of choice saliently explain why she chooses one action over another by influencing what reasons come to mind and how strong those reasons are. Such actions are circumstantially lucky. It is also possible that a person's reason-giving dispositions do decisively support one action and that those dispositions are not constitutively lucky, because they were previously modified by circumstantially lucky actions; but this possibility also is not a luck-free avenue for self-creation. Thus, because history-sensitive compatibilists agree that constitutive luck negates freedom and there is no luck-free way to undo constitutive luck, history-sensitive compatibilists should agree that we cannot act freely. That, in brief, is the Luck Pincer.

Galen Strawson's (1986, pp. 21–51, 1994) Basic Argument purports to show that it is impossible to be morally responsible for anything entirely due to constitutive luck. In Hartman (2018, pp. 165–166), I offer this summary:

Reasons Premise: An agent *S*'s intentionally performing an action *A* for which she might be morally responsible is explained by certain features of her mental constitution *MC* – namely, certain reasons for acting.

Responsibility Premise: S is morally responsible for an intentional action A only if S is morally responsible for the parts of her MC that explain her performing A, and S is morally responsible for her MC only if S is morally responsible for an earlier action A1 in which S intentionally and successfully brought about those parts of her MC.

Iteration Premise: S is morally responsible for A by way of MC and A1 as previously described only if S has performed an infinite number of even earlier free actions. (After all, S is morally responsible for A1 only if S is morally responsible for the parts of her MC1 that explain her performing A1, and S is morally responsible for her MC1 only if S is morally responsible for an even earlier action A2 in which S intentionally and successfully brought about those parts of her MC1. S is morally responsible for A2 only if ...).

Impossibility Premise: It is impossible for finite beings like us to have performed an infinite number of past actions.

Conclusion: It is impossible for finite beings like us to be morally responsible for anything.

So, because a person must be morally responsible at least to some degree for the mental dispositions that explain an action to be morally responsible for that action and because a person's first action must be performed from mental dispositions that are entirely constitutively lucky, it is impossible to be morally responsible for that first choice or for any choices thereafter.

Is the Basic Argument sound? The argument is valid. The only premise that might plausibly be denied is the Responsibility Premise (see Hartman 2018, p. 166). So, whether a person finds the Basic Argument compelling depends on whether they find the Responsibility Premise to be plausible. Numerous philosophers think that the Responsibility Premise is exactly right (see, for example, Istvan 2011; Kershnar 2015). Many others reject the Responsibility Premise as too demanding to describe plausibly a necessary condition on acting freely (see, for example, Clarke 2005; Fischer 2006; Hartman 2018).

In my view, Strawson's conception of moral responsibility is too demanding. It is not the case that a person's being morally responsible for something requires an infinity of choices for which she is morally responsible or an incoherent state of affairs in which she makes her first free choice from previously freely chosen character. Rather, moral responsibility for choices just emerges from the right kind of non-responsibility conditions (Hartman 2018; see also Mele 2006, pp. 129–133; Cyr 2019). Moral responsibility for action emerges when a person has the right kind of knowledge of morality and mundane matters of fact and has the capacities that grant her freedom-relevant control over her action, which may be accounted for, for example, in terms of being properly responsive to reasons (Fischer and Ravizza 1998) or having the agent-causal power to choose between various possible actions (O'Connor 2000).

Rejection of the Responsibility Premise has a constructive implication for the nature of free will, because it is a crucial part in a broader argument by Taylor Cyr (2020) against the previously mentioned history-sensitivity condition on directly free actions (see also Kane 2007, pp. 174–175; Lemos 2018, pp. 34–35).

The denial of the Responsibility Premise implies that, for example, a youth's not being morally responsible for her character does not itself rule out the possibility of her performing her first free action. Cyr (2020, pp. 2387–2391) argues that there is no relevant difference with respect to being morally responsible to some extent for an action between a youth who is not morally responsible to any extent for her character and Vic, our victim of neuroscientific transformation, who is not morally responsible to any extent for his vices. After all, the character traits that motivate both of their actions are constitutively lucky. For the sake of

argument, assume that the youth and Vic have all the relevant knowledge and capacities that compatibilists or libertarians think are sufficient for acting freely, for example, being reasons-responsive or having multi-track agent-causal power. In that case, if the youth with those capacities can be morally responsible to some degree for her first free action influenced by her constitutively lucky dispositions, then the same should be true of Vic for his free action influenced by his constitutively lucky dispositions, because Vic has the same basic capacities required for free action but has them to a greater extent.¹²

If Vic is morally responsible for his action at least to some degree, history-sensitive conditions on acting freely turn out to be false,¹³ because these conditions are proposed to ensure that a person in Vic's condition is not morally responsible to any degree for his action. Thus, compatibilists such as Fischer and Ravizza (1998, pp. 170–239), Haji and Cuypers (2007), and Mele (2019) who reject the Responsibility Premise but affirm various history-sensitive conditions should reject those history-sensitive conditions.¹⁴

Let us return to the Luck Pincer. Cyr's argument and the rejection of the Responsibility Premise provide a joint response to the Luck Pincer – namely, we should reject the history-sensitive requirement on directly free actions. Why is this a response to the Luck Pincer? The history-sensitive requirement is what justifies its premise that constitutive luck must be undone prior to acting freely, and so by rejecting that requirement, the Luck Pincer cannot get off the ground without a new argument for the premise that constitutive luck must be undone prior to acting freely. This response imposes a dialectical hardship on proponents of the Luck Pincer, because it is difficult to provide such an argument. Levy (2019, p. 67) himself writes, “I confess I have no response to the question what it is about luck that undermines moral responsibility other than to point to clear cases and common intuitions ... Here we hit ground level.” This is a common refrain; for example, David Enoch (2010, p. 30) writes, “It is very hard to argue for – or, indeed, against – moral luck, because – as is often noted – the necessary conclusions for either position are extremely close to the relevant premises ... [and close to] moral bedrock.” Thus, before the Luck Pincer gives us a good reason to accept its conclusion, we need a new argument for the premise that constitutive luck must be undone prior to acting freely.

Furthermore, we have reason to reject Levy's claim that luck undermines free will. Levy (2011, p. 1) himself characterizes free will as the control condition on moral responsibility, as I did earlier in this chapter; yet Levy (2019, p. 67) grants the conclusion of my arguments in Hartman (2017, pp. 51–55) that luck does not undermine moral responsibility *by* undermining control: “Constitutive and circumstantial luck do not reduce our control over our actions; they threaten our responsibility in other ways” (Levy 2019, p. 66; cf. Caruso 2019). But if constitutive and circumstantial luck in general do not mitigate control and if a sufficient condition for free will is characterized in terms of control, constitutive and circumstantial luck in general do not negate free will.

This unified response to the Basic Argument and the Luck Pincer motivates exploring the relationships between compatibilism and libertarianism, on the one hand, and various kinds of moral luck, on the other.

4 Compatibilism and Moral Luck

Recall that compatibilism is the view that acting freely is not ruled out just because the action is causally determined by factors beyond the actor's control. Add to this definition of compatibilism that human persons at least sometimes act freely.¹⁵

Compatibilism implies the possibility of causal moral luck. *Causal moral luck* occurs when the laws of nature and past states of affairs outside of a person's control causally determine her actions and thereby positively affect her degree of moral responsibility.¹⁶ Compatibilism implies that causal luck does not necessarily negate moral responsibility; so, causal luck can positively affect her degree of moral responsibility. Whether causal moral luck exists depends on whether the actions for which people are morally responsible are causally determined by factors beyond their control.¹⁷

Does compatibilism imply that at least some kinds of constitutive and circumstantial moral luck exist? Let us first make the relevant definitions explicit. *Circumstantial moral luck* occurs when a person faces a morally significant challenge that is outside of her control, and it affects her positive praiseworthiness or blameworthiness. In concrete terms, Merely Reckless's being more blameworthy than No Start is a case of circumstantial moral luck, because their different circumstantial luck makes a difference to their degrees of blameworthiness. *Constitutive moral luck* occurs when an agent possesses dispositions or capacities in a way that is outside of her control, and they affect her positive praiseworthiness or blameworthiness for a trait or an action. In concrete terms, Merely Reckless's being more blameworthy than Scarred Childhood is a case of constitutive moral luck, because their different constitutive luck makes a difference to their degrees of blameworthiness.

In Hartman (2017, pp. 53–54), I argued that extant causal moral luck does imply extant constitutive and circumstantial moral luck, but Levy's (2019, p. 66) response convinced me that those arguments overgeneralize in unacceptable ways. Here I sketch a more modest version of the argument that does not unacceptably overgeneralize.

The existence of causal moral luck provides some evidence for the existence of constitutive and circumstantial moral luck. Compatibilism implies that an action's being causally lucky in general yields no excuse. If causal luck were to excuse, it would excuse due to force, because causal luck has such great influence on an action that it guarantees that it occurs in the way that it does. The influence of mere constitutive and circumstantial luck in general is not so great, because these lucky features are merely one among many other influences on our actions. A lucky constitution does not guarantee the occurrence of an action or omission absent a circumstance, and a lucky circumstance does not guarantee the occurrence of an action or omission absent mental dispositions. So then, because the greater influence of causal luck in general does not excuse, this provides evidence that the same is true of constitutive and circumstantial luck in general, which have less influence on a person's actions and omissions.¹⁸ So, since we have evidence that constitutive and circumstantial luck in general do not undermine freedom, we have evidence that such luck can positively affect a person's moral responsibility for actions and omissions.

Mainstream compatibilists accept at least some kinds of circumstantial and constitutive moral luck. On the one hand, some *quality of will* compatibilists hold the view that a person is blameworthy for a character trait if it is bad and sensitive to judgment, regardless of whether the trait is part of her genetic endowment or is produced in her by neuroscientific indoctrination (see Smith 2005). Such views embrace a large amount of constitutive moral luck. One might think that this view implies that circumstantial moral luck cannot exist, because, for example, No Start and Merely Reckless are equally blameworthy due to their identical bad character traits (see Talbert 2019, pp. 34–35). But even this view is committed at least to a *diachronic* kind of circumstantial moral luck. Different lucky circumstantial opportunities and challenges prompt different kinds of actions and thereby prompt different kinds of character formation (Hartman 2020a, p. 108); in concrete terms, the unique actions of Merely Reckless can have a unique impact on her character through time and can thereby differentiate her

degree of blameworthiness from No Start through time. On the other hand, *reasons-responsive* compatibilists embrace far less constitutive moral luck. They regard persons as moral responsibility for traits if and only if they are formed in the right kind of way by their past directly free actions (see Fischer and Ravizza 1998; Fischer and Tognazzini 2009). And they embrace far more circumstantial moral luck. On such views, persons are fundamentally morally responsible for the exercises of their reasons-responsive capacities, and at least most opportunities to exercise those capacities are shaped by circumstantial luck. Thus, mainstream compatibilists do accept causal, constitutive, and circumstantial moral luck.

Should such compatibilists accept resultant moral luck? *Resultant moral luck* occurs when the consequence of a person's action is at least partially beyond her control, and the consequence partially determines her positive praiseworthiness or blameworthiness; in concrete terms, Killer's being more blameworthy than Merely Reckless is a case of resultant moral luck, because their different resultant luck makes a difference to their degrees of blameworthiness.

Michael Moore (1997, pp. 233–246) argues that compatibilists should accept resultant moral luck in a way that may be summarized as follows (cf. Russell 2017a):

- (1) If resultant moral luck cannot exist, then constitutive and circumstantial moral luck cannot exist either.
 - (2) Constitutive and circumstantial moral luck can exist.
- Thus,
- (3) Resultant moral luck can exist too.

I argued that compatibilism provides some evidence for (2) and described mainstream compatibilists as accepting (2). What can be said for (1)? Moore (1997, p. 237) justifies (1) by appealing to this consideration: "luck is luck, and to the extent that causal fortuitousness is morally irrelevant anywhere it is morally irrelevant everywhere."

But there are problems with (1). First, Moore's justification appears merely to restate the claim that it is supposed to justify. Second, it is not obvious that (1) is true. After all, one might be sympathetic to the allegedly Kantian view that we are morally responsible only for our inner states of willing and so not for the consequences of our actions (for example, Khoury 2018);¹⁹ in terms of our example, Killer is not even morally responsible for killing the pedestrian. Alternatively, one might think that we are morally responsible for the consequences of our actions, but that such consequences do not affect our degree of moral responsibility (for example, Thomson 1989, pp. 208–211; cf. Zimmerman 2002, pp. 560–561); in concrete terms, Killer is morally responsible for killing a pedestrian whereas Merely Reckless is not, but Killer does not thereby deserve more blame than Merely Reckless. They are equally blameworthy.

These options highlight ways in which compatibilists can resist Moore's argument for resultant moral luck. In the final section, I point to a more plausible reason for the claim that compatibilists should accept resultant moral luck.

5 Libertarianism and Moral Luck

Libertarianism is the view that a person's action being a free action is incompatible with the action's being causally determined by factors beyond her control and that human persons act freely at least sometimes.

Libertarianism implies that causal moral luck cannot exist. If acting freely is ruled out by causal luck, causal luck cannot *positively* affect a person's degree of moral responsibility for an action.

Does libertarianism also imply that constitutive, circumstantial, and resultant moral luck cannot exist? It depends on the motivations for libertarianism.

One motivation for libertarianism is the idea that causal moral luck cannot exist, because all luck must be factored out of moral responsibility (cf. Russell 2017b, pp. 145–149). Recall that this is precisely the motivation driving skeptical arguments from luck. As a result, for these libertarians to secure their commitment that human persons act freely at least sometimes, they tend to embrace extreme metaphysical commitments. Consider two examples.

First, there is the Kantian view that our noumenal free agency floats free of the influence of circumstantial and constitutive luck (see Athanassoulis 2005, pp. 104, 112).

This description of free choice seems implausible considering the ways in which our constitutive and circumstantial luck influence our apparent free choices. Even if this objection could be surmounted by embracing the view that free choices are rare (Campbell 1951, pp. 460–461), if free choices are entirely uninfluenced by constitutive and circumstantial luck, such choices would be wholly inexplicable and arbitrary due to the complete absence of a rationalizing explanation, which poses a problem for their being free choices (Hartman 2017, p. 98).²⁰

Second, there is Michael Zimmerman's (2002) counterfactual view. This view neutralizes resultant moral luck, because consequences do not affect a person's degree of moral responsibility; in terms of our example, Killer and Merely Reckless are equally blameworthy. This view also neutralizes at least most circumstantial and constitutive moral luck, because persons are praiseworthy and blameworthy not only for what they actually freely do but also in virtue of what they would have freely done in counterfactual circumstances, including modally distant counterfactual circumstances with counterfactual character and history.²¹ In concrete terms, Killer, Merely Reckless, No Start, and Scarred Childhood are equally blameworthy, because they would do the same thing in one another's lucky circumstances and with one another's lucky constitutions.

Zimmerman's view is difficult to accept at least for three reasons. First, it requires that there be infinitely many true counterfactuals of libertarian freedom to rule out moral luck, because there is an infinite number of circumstances in which a person could possibly be. But there is a good reason for libertarians to deny that there are any such true counterfactuals (Hartman 2017, pp. 71–73). Suppose that freedom requires alternative possibilities at the moment of choice. In that case, nothing about her actual states of character, mind, or circumstance can make it true that she would perform or omit a free action until she actually exercises her free agency. But because she cannot *actually* exercise her free agency in a *counterfactual* circumstance, such counterfactuals cannot be true.²² Second, even if there are true counterfactuals of freedom, Zimmerman's view cannot fulfill its own luck-free aspiration if a person has some of her lucky constitutive properties essentially, because it is metaphysically impossible for a person to be in a circumstance with different essential constitutive luck (Hartman 2019c, pp. 3184–3187). Third, if there are satisfactory answers to these objections, Zimmerman's view still counterintuitively implies that people are infinitely praiseworthy and blameworthy in virtue of the relevant true counterfactuals, because it is plausible that praiseworthiness and blameworthiness do not cancel out each other (Hartman 2017, p. 66).

The most viable forms of libertarianism have a different motivation for denying causal moral luck. Such libertarians offer a reason why causal luck is pernicious to

freedom that does not apply generally to the other kinds of luck. For example, causal luck necessarily precludes our being the source of our actions in a way that is essential to acting freely, but constitutive and circumstantial luck do not (Pereboom 2014, pp. 9–29, 74–82). Or causal luck necessarily rules out the kind of alternative possibilities at the moment of choice that are essential to acting freely, but constitutive and circumstantial luck do not (van Inwagen 1983, pp. 55–105). To put the latter point in metaphorical terms, freedom requires that our world be a garden of forking paths such that the free agent has the power to choose which path to take and thereby the power to take one path or the other. Causal luck necessarily rules out the possibility of more than one open path, but constitutive and circumstantial luck often do not rule out the possibility of more than one open path; and in cases in which constitutive or circumstantial luck do so, they are cases of causal luck rather than mere constitutive or circumstantial luck. Thus, the most viable motivation for libertarianism does not rule out the possibility of constitutive and circumstantial moral luck. Such libertarians tend also to embrace at least some constitutive and circumstantial moral luck, but how much they embrace depends on particularities of the accounts.

6 Compatibilism, Libertarianism, and Resultant Moral Luck

Compatibilists and libertarians who accept the unified response to the Luck Pincer and Basic Argument and three additional claims yet to be elaborated have a reason to accept the existence of resultant moral luck.

The unified response is that a person can be morally responsible to some degree for an action that is entirely motivated by constitutively lucky dispositions if such an action satisfies the relevant compatibilist or libertarian conditions on moral responsibility. In concrete terms, Vic can be morally responsible to some degree for his free choice to murder.

Consider the three additional claims. First, Vic is not very morally responsible or blameworthy for his free action due to his bad constitutive luck. Thus, acting from a purely constitutively lucky character can be a *partial* excuse, because a person's degree of moral responsibility is mitigated, but not eliminated, by an action motivated purely by constitutively lucky dispositions. This accommodates a kernel of truth of the Responsibility Premise that being morally responsible for the way you are is important in some way to being morally responsible for an action.

Second, Vic is much less blameworthy for his action than someone – call him Vic* – who is blameworthy to a significant extent for the character that motivates and explains his type-identical murder.²³ What this judgment reveals is that being morally responsible for the character that motivates and explains an action is relevant to the agent's degree of moral responsibility for the action (O'Connor 2005, pp. 219–220; Kane 2007, pp. 174–175; Rogers 2015, pp. 127–150; Lemos 2018, pp. 34–35, 68–73; Cyr 2020, pp. 2392–2393; Hartman 2020b, pp. 1422–1424). In my view, being morally responsible to a significant extent for the character that motivates and explains an action is required to be *fully* morally responsible for the action (Hartman manuscript).

Third, a person is morally responsible for a character trait to some extent if and only if the trait was generated, strengthened, weakened, or maintained by actions or omissions for which she is directly morally responsible and that make a difference to her character in a way that she could reasonably have been expected to foresee (Hartman 2020b, pp. 1426–1431; cf. Fischer and Tognazzini 2009).

Compatibilists and libertarians who embrace the unified response and these three plausible claims have a reason to accept the existence of resultant moral luck (see also Cyr 2021, pp. 306–307). Why think that? Moral responsibility for character is a species of moral responsibility for consequences, which gives us the “resultant” part of resultant moral luck. Moral responsibility for consequences can affect a person’s degree of moral responsibility, because moral responsibility for character can increase her degree of moral responsibility for the actions explained and motivated by that character, which gives us the “moral” part of resultant moral luck. Consequences that affect character are lucky, which gives us the “luck” part.

The “resultant” and “moral” parts easily fall out of the general view. But why is the “luck” part true? That is, why are character-affecting consequences lucky? It is perhaps easiest to see why via example and generalization. Suppose that Jan wants to grow in temperance. Jan freely removes all the junk food from her home; one foreseeable consequence of that free choice is that she no longer has the option to choose junk food at home, because her choices at home are restricted by what is easily available. Her temperance strengthens as a result. In particular, she gains new beliefs about the benefits of eating healthy such as having more energy, and she gains new desires for healthy food by habituation. But this character forming consequence is subject to luck. After all, Jim is Jan’s spouse, and Jim’s behavior is outside of Jan’s control; Jim might reintroduce junk food by bringing home Flamin’ Hot Limón Doritos. Such an event would diachronically alter the consequence of having a junk food-free home, which illustrates the way in which that consequence is lucky as well as the way in which the consequence helps to strengthen her temperance is lucky. But then, Jan’s free choice to influence her character is subject to resultant luck, and so is the way in which she is morally responsible for the changes to her character influenced by that free choice. The case generalizes, because character forming strategies are almost always indirect in this way,²⁴ and so the foreseeable actual consequences of character shaping free actions are almost always lucky.

Thus, compatibilists and libertarians have a reason to embrace resultant moral luck in view of the unified response and those three claims.^{25,26}

Notes

- 1 For recent exceptions, see Cyr (2019, 2020), Ekstrom (2019), Hartman (2018, 2020b), Mickelson (2015, 2019), Pérez de Calleja (2014, 2019), Swenson (2019), and Tognazzini (2011).
- 2 The nature of the moral responsibility at the heart of the free will and moral luck debates is what Pereboom (2014, p. 2) calls “basic desert” moral responsibility, according to which “the agent would deserve to be blamed or praised just because she has performed the action ... and not, for example, merely by virtue of consequentialist or contractualist considerations.” According to Shoemaker’s (2015) pluralism, the kind of moral responsibility at stake is accountability rather than attributability or answerability.
- 3 What is the incompatibilist relation? Is it a *metaphysical incompatibility* relation, according to which one relatum rules out the other, or is it an *impossibility* relation, according to which the relata merely cannot obtain together (see Mickelson, this volume)? I have in mind the former relation, but readers who prefer the latter may make the relevant changes.
- 4 Nagel’s own term is “antecedent luck”. Some philosophers broaden the category of causal luck to include deterministic *and* indeterministic luck (see Mickelson 2019, p. 225, no. 3; Sartorio 2019, p. 211).
- 5 Nagel (1979, p. 34) embraces a “paradox” by accepting this conclusion and that we are morally responsible for at least some of our actions.

- 6 In my view, Michael Zimmerman (1987) refutes Nagel's (1979) skeptical argument. John Greco (1995) develops a better version of Nagel's skeptical worry, but Greco's argument is too complicated to summarize here; ultimately, Greco argues against the skeptical argument that he develops (cf. Hartman 2017, pp. 124–139).
- 7 See Hartman (2017, p. 23) for citations prior to 2017.
- 8 Again, see Hartman (2017, p. 23) for a more expansive citation list.
- 9 For more arguments that support the standard definition, see Anderson (2019), Hartman (2017, pp. 23–31), and Statman (2019); for a new challenge, see Levy (2019).
- 10 The precise reason why Vic is not blameworthy differs for different versions of the history-sensitive condition.
- 11 Levy (2011, pp. 11–40, 2019) offers a more complex account of luck; I omit those details because my unified reply does not hinge on them.
- 12 I return to what else we should say about Vic in the final section.
- 13 Recall that the relevant history-sensitive conditions are necessary conditions on *directly* free actions. The falsity of such conditions is compatible with there being history-sensitive conditions on *indirect* moral responsibility. For example, a drunk driver could still be indirectly morally responsible for killing the pedestrian due to past directly free choices to get drunk and directly free omissions to make plans to get home safely.
- 14 Mele (1995, 2006) advances both compatibilist and libertarian accounts of free agency, and he proposes history-sensitive conditions for both.
- 15 The addition makes this definition symmetrical to a standard definition of libertarianism.
- 16 If the reader does not think that moral responsibility comes in degrees, she can substitute “degrees of moral praiseworthiness and blameworthiness”.
- 17 Most compatibilists think that freedom is also compatible with indeterminism (see Fischer 2011).
- 18 Certain kinds of causal, constitutive, and circumstantial luck do excuse (see Hartman 2017, pp. 90–95; 2018, pp. 176–180).
- 19 In Hartman (2019b), I argue that Kant does not actually hold this view.
- 20 This objection is different from the standard luck objection to libertarianism, because the standard objection does not describe an agent's making a choice without dispositional anchors.
- 21 Zimmerman (2002, p. 573) advertises his luck-free strategy as equally available to libertarians *and* compatibilists. But in Hartman (2017, pp. 73–75), I argue that it is unmotivated for anyone who embraces Zimmerman's view to affirm causal moral luck.
- 22 If a person is morally responsible for character due to its having been formed by past free actions and that character determines what she would do in some counterfactual circumstance, there can be true counterfactuals of *indirect* freedom (see Hartman 2017, pp. 78–80; 2020b, pp. 1419–1420). But such counterfactuals cannot offer much aid to Zimmerman's view. Actual character grounds true counterfactuals only when actual dispositions are *kept fixed* in the antecedent. But to eliminate constitutive luck, a person must have been in the formative circumstances of others, which requires that actual dispositions are *not kept fixed*.
- 23 One might be tempted to respond that Vic* is morally responsible for more things – namely, his character and action – rather than being more morally responsible (McKenna 2004, p. 183). But I do not think that this correctly identifies the moral properties of the actions (Hartman 2020b, p. 1424, no. 11).
- 24 Miller's (2017, pp. 169–254) list of indirect strategies include going to therapy, surrounding ourselves with virtuous people, setting goals, virtue-labelling ourselves, nudging ourselves, selecting our circumstances, joining a religious community, and asking God for help.
- 25 In Hartman (2017, pp. 105–111, 124–138; 2019c, pp. 3188–3192), I argue that compatibilists and libertarians have yet other reasons to embrace extant resultant moral luck.
- 26 I thank Olle Blomberg, Taylor Cyr, Matt King, Neil Levy, Kristin Mickelson, Jennifer Page, and Abelard Podgorski for comments on a draft of this paper.

Bibliography

- Anderson, M.B. (2019). Moral luck as moral lack of control. *Southern Journal of Philosophy* 57: 5–29. doi: 10.1111/sjp.12317.
- Athanassoulis, N. (2005). *Morality, Moral Luck, and Responsibility: Fortune's Web*. New York: Palgrave Macmillan.
- Campbell, C.A. (1951). Is 'freewill' a pseudo-problem? *Mind* 60: 441–465. doi: 10.1093/mind/103.409.35.
- Caruso, G.D. (2019). In defense of the luck pincer: why luck (still) undermines moral responsibility. *Journal of Information Ethics* 28: 51–72.
- Church, I.M. and Hartman, R.J. eds. (2019). Luck: an introduction. In: *The Routledge Handbook of the Philosophy and Psychology of Luck*, 1–10. New York: Routledge.
- Clarke, R. (2005). On an argument for the impossibility of moral responsibility. *Midwest Studies in Philosophy* 29: 13–24. doi: 10.1111/j.1475-4975.2005.00103.x.
- Cyr, T.W. (2019). Moral responsibility, luck, and compatibilism. *Erkenntnis* 84: 193–214. doi: 10.1007/s10670-017-9954-7.
- Cyr, T.W. (2020). Manipulation and constitutive luck. *Philosophical Studies* 177: 2381–2394. doi: 10.1007/s11098-019-01315-y.
- Cyr, T.W. (2021). The inescapability of moral luck. *Thought* 10: 302–310. doi: 10.1002/tht3.508.
- Ekstrom, L.W. (2019). Luck and libertarianism. In: *The Routledge Handbook of the Philosophy and Psychology of Luck* (ed. I.M. Church and R.J. Hartman), 239–247. New York: Routledge.
- Enoch, D. (2010). Moralluck and the law. *Philosophy Compass* 5: 42–54. doi: 10.1111/j.1747-9991.2009.00265.x.
- Fischer, J.M. (2006). The cards that are dealt you. *Journal of Ethics* 10: 107–129. doi: 10.1007/s10892-005-4594-6.
- Fischer, J.M., ed. (2011). Indeterminism and control: an approach to the problem of luck. In: *Deep Control: Essays on Free Will and Value*, 85–105. New York: Oxford University Press.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
- Fischer, J.M. and Tognazzini, N.A. (2009). The truth about tracing. *Noûs* 43: 531–556. doi: 10.1111/j.1468-0068.2009.00717.x.
- Greco, J. (1995). A second paradox concerning responsibility and luck. *Metaphilosophy* 26: 81–96. doi: 10.1111/j.1467-9973.1995.tb00557.x.
- Haji, I. and Cuypers, S. (2007). Magical agents, global induction, and the internal/externalism debate. *Australasian Journal of Philosophy* 85: 343–371. doi: 10.1080/00048400701571602.
- Hartman, R.J. (2017). *In Defense of Moral Luck: Why Luck Often Affects Praiseworthiness and Blameworthiness*. New York: Routledge.
- Hartman, R.J. (2018). Constitutive moral luck and Strawson's argument for the impossibility of moral responsibility. *Journal of the American Philosophical Association* 4: 165–183. doi: 10.1017/apa.2018.18.
- Hartman, R.J. (2019a). Accepting moral luck. In: *The Routledge Handbook of the Philosophy and Psychology of Luck* (ed. I.M. Church and R.J. Hartman), 227–238. New York: Routledge.
- Hartman, R.J. (2019b). Kant does not deny resultant moral luck. *Midwest Studies in Philosophy* 43: 136–150. doi: 10.1111/misp.12109.
- Hartman, R.J. (2019c). Moral luck and the unfairness of morality. *Philosophical Studies* 176: 3179–3197. doi: 10.1007/s11098-018-1169-5.
- Hartman, R.J. (2020a). Against the character solution to the problem of moral luck. *Australasian Journal of Philosophy* 98: 105–118. doi: 10.1080/00048402.2019.1617755.
- Hartman, R.J. (2020b). Indirectly free actions, libertarianism, and resultant moral luck. *Erkenntnis* 85: 1417–1436. doi: 10.1007/s10670-018-0084-7.
- Hartman, R.J. Manuscript. *Character and Free Will*.

- Istvan Jr., M.A. (2011). Concerning the resilience of Galen Strawson's basic argument. *Philosophical Studies* 155: 399–420. doi: 10.1007/s11098-010-9578-0.
- Kane, R. (2007). Response to Fischer, Pereboom, and Vargas. In: *Four Views on Free Will* (ed. J.M. Fischer, R. Kane, D. Pereboom and M. Vargas), 166–184. Malden: Blackwell Publishing.
- Kershnar, S. (2015). Moral responsibility and foundationalism. *Philosophia* 43: 381–402. doi: 10.1007/s11406-015-9586-6.
- Khoury, A.C. (2018). The objects of moral responsibility. *Philosophical Studies* 175: 1357–1381. doi: 10.1007/s11098-017-0914-5.
- Latus, A. (2003). Constitutive luck. *Metaphilosophy* 34: 460–475. doi: 10.1111/1467-9973.00285.
- Lemos, J. (2018). *A Pragmatic Approach to Libertarian Free Will*. New York: Routledge.
- Levy, N. (2011). *Hard Luck: How Luck Undermines Freedom and Moral Responsibility*. Oxford: Oxford University Press.
- Levy, N. (2019). Putting the luck back in moral luck. *Midwest Studies in Philosophy* 43: 59–74. doi: 10.1111/misp.12104.
- McKenna, M. (2004). Responsibility and globally manipulated agents. *Philosophical Topics* 32: 169–192. doi: 10.5840/philtopics2004321/222.
- Mele, A.R. (1995). *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press.
- Mele, A.R. (2006). *Free Will and Luck*. New York: Oxford University Press.
- Mele, A.R. (2019). *Manipulated Agents: A Window into Moral Responsibility*. New York: Oxford University Press.
- Mickelson, K. (2015). The zygote argument is invalid: now what? *Philosophical Studies* 172: 2911–2929. doi: 10.1007/s11098-015-0449-6.
- Mickelson, K.M. (2019). Free will, self-creation, and the paradox of moral luck. *Midwest Studies in Philosophy* 43: 224–256. doi: 10.1111/misp.12114.
- Miller, C.B. (2017). *The Character Gap: How Good Are We?* New York: Oxford University Press.
- Moore, M.S. (1997). *Placing Blame: A General Theory of Criminal Law*. New York: Oxford University Press.
- Nagel, T., ed. (1979). Moral luck. In: *Mortal Questions*, 24–38. New York: Cambridge University Press.
- O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.
- O'Connor, T. (2005). Freedom with a human face. *Midwest Studies in Philosophy* 29: 207–227. doi: 10.1111/j.1475-4975.2005.00113.x.
- O'Connor, T. and Franklin, C. (2018). Free will. *The Stanford Encyclopedia of Philosophy* (ed. N.Z. Edward). <https://plato.stanford.edu/entries/freewill>.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.
- Pérez de Calleja, M. (2014). Cross-world luck at the time of decision is a problem for compatibilists as well. *Philosophical Explorations* 17: 112–125. doi: 10.1080/13869795.2014.912673.
- Pérez de Calleja, M. (2019). Luck and compatibilism. In: *The Routledge Handbook of the Philosophy and Psychology of Luck* (ed. I.M. Church and R.J. Hartman), 248–258. New York: Routledge.
- Rogers, K.A. (2015). *Freedom and Self-Creation: Anselmian Libertarianism*. New York: Oxford University Press.
- Russell, P., ed. (2017a). Free will pessimism. In: *The Limits of Free Will*, 243–276. New York: Oxford University Press.
- Russell, P., ed. (2017b). Free will, art, and morality. In: *The Limits of Free Will*, 134–158. New York: Oxford University Press.
- Sartorio, C. (2019). Kinds of moral luck. In: *The Routledge Handbook of the Philosophy and Psychology of Luck* (ed. I.M. Church and R.J. Hartman), 206–215. New York: Routledge.
- Shoemaker, D. (2015). *Responsibility from the Margins*. New York: Oxford University Press.
- Smilansky, S. (2000). *Free Will and Illusion*. New York: Oxford University Press.
- Smith, A.M. (2005). Responsibility for attitudes: activity and passivity in mental life. *Ethics* 115: 236–271. doi: 10.1086/426957.
- Statman, D. (2019). The definition of luck and the problem of moral luck. In: *The Routledge Handbook of the Philosophy and Psychology of Luck* (ed. I.M. Church and R.J. Hartman), 195–205. New York: Routledge.

- Strawson, G. (1986). *Freedom and Belief*, 2e. Oxford:Oxford University Press.
- Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies* 75: 5–24. doi: 10.1007/BF00989879.
- Swenson, P. (2019). Luckily, we are only responsible for what we could have avoided. *Midwest Studies in Philosophy* 43: 106–118. doi: 10.1111/misp.12107.
- Talbert, M. (2019). The attributionist approach to moral luck. *Midwest Studies in Philosophy* 43: 24–41. doi: 10.1111/misp.12102.
- Thomson, J.J. (1989). Morality and bad luck. *Metaphilosophy* 20: 203–221. doi: 10.1111/j.1467-9973.1989.tb00423.x.
- Tognazzini, N.A. (2011). Owning up to luck. *Social Theory and Practice* 37: 95–112. doi: 10.5840/soctheorpract20113717.
- van Inwagen, P. (1983). *An Essay on Free Will*. New York: Oxford University Press.
- van Inwagen, P. (2008). How to think about the problem of free will. *The Journal of Ethics* 12: 327–341. doi: 10.1007/s10892-008-9038-7.
- Watson, G. (1987). Responsibility and the limits of evil: variations on a Strawsonian theme. In: *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (ed. F. Schoeman), 256–286. Cambridge: Cambridge University Press.
- Wolf, S. (1990). *Freedom within Reason*. New York: Oxford University Press.
- Zimmerman, M.J. (1987). Luck and moral responsibility. *Ethics* 97: 374–386. doi: 10.1086/292845.
- Zimmerman, M.J. (2002). Taking luck seriously. *The Journal of Philosophy* 99: 553–576. doi: 10.2307/3655750.

Basic Desert and the Appropriateness of Blame

KELLY MCCORMICK

Consider a case in which all of your colleagues have a strong preference for using a certain kind of pen. Unfortunately, due to budgetary constraints, only one box of this vastly superior kind of pen can be included in your department's monthly order of office supplies, and there are only enough in each box for each member of the department to take one. In order to promote collegiately, you have all agreed to refrain from taking an extra pen. Alas, one of your colleagues decides one day to abandon this agreement and, in secret, sneaks into the department and takes the entire new box of vastly superior pens. To make matters worse, you arrive to your office the next day and discover not only the unfortunate pen shortage, but also that the thief has told all of your colleagues that she saw you leaving the department with the missing pens the previous evening. For whatever reason, your colleagues believe the actual thief (perhaps you have an unfortunate history of sneaking more than your fair share of office supplies), and they are angry. Further, they believe you have knowingly and intentionally done something *wrong* in taking the pens, and thus their anger has a particular moral flavor. They resent and blame you for the shortage of vastly superior pens they must all now suffer.

Do you deserve to be the object of your colleagues' blame in this case? Of course not. You've done nothing wrong, and wrongdoing looks to be an obvious necessary condition¹ for desert of blame. But paradigmatic examples of failing to deserve blame are cheap. It is far more difficult to specify the conditions under which an agent is deserving of blame, is deserving of blame *in the basic sense*, and furthermore why this is so.

To make this latter notion clearer consider two variations on the above case. In both variations you are in fact the pen thief. In the first variation, assume that some success theory² of moral responsibility is correct, and when you steal the pens you satisfy all of the necessary and sufficient conditions for responsibility according to the correct theory. Your colleagues recognize this, and blame and resent you for the pen shortage *only because you satisfy these conditions*. In the second variation, assume that some brand of eliminativism³ about moral responsibility is correct. Again, both you and your colleagues are aware of this fact. But, the administration at your university has recently implemented the following policy in order to crack down on rampant misuse of office supplies: whenever an employee takes home office supplies not

allocated to them, their colleagues must resent them or their home department will lose a tenure line. In light of this your colleagues also resent you in the second variation, but *only because of the policy*. While there may be some sense in which you deserve blame in the second variation of the case – you are, after all, aware of the policy and the fact that you are subject to it like everyone else – it is intuitively not the basic sense. And while the first variation looks like a better candidate for an example of basic desert, it is far from obvious why this is so.

In fact, it is notoriously difficult to articulate what basic desert of blame amounts to. This notion appears to be, as Michael McKenna has recently put the point, “the elephant in the room” in discussions of free will, moral responsibility, and blame (2012, p. 120). However, appeals to basic desert (or lack thereof) are often wielded by philosophers in these debates. For example, one way of understanding the force of Derk Pereboom’s Four Case Manipulation Argument is that it shows that determinism is incompatible with basic desert of moral praise and blame.⁴ Thus any brand of compatibilism will be unsatisfying if the agents who meet the proposed compatibilist conditions for responsibility still fail to deserve moral praise and blame in the basic sense.⁵ Another instructive example is the role that basic desert plays in recent exchanges between eliminativists about moral responsibility, like Pereboom, and revisionists like Manuel Vargas.⁶ One question central to the success of revisionism is whether or not this kind of view can justify our responsibility-related practices, in particular those that involve moral praising and blaming. These practices, as Pereboom and others claim,⁷ presuppose that agents do in fact sometimes deserve praise and blame in the *basic* sense. And so revisionists, like compatibilists, must show that the account of responsibility they offer captures a form of agency capable of delivering basic desert, or they must justify further revisionism about the normative grounds for our blaming practices.⁸

The goal of this paper is to attempt to capture the shape of the elephant, albeit roughly, and in a way that does not stack the deck against any of the particular disputants in these debates. I begin in section 1 by identifying a number of relatively uncontroversial features that serve to provide a first step in distinguishing between basic and non-basic desert. Unfortunately these features provide only a negative characterization of basic desert, but restricting focus to basic desert of blame in particular suggests some further positive features. In section 2 I turn to a particular attempt to offer a positive account of what makes agents deserving of blame offered recently by Michael McKenna (2012). McKenna characterizes this account as a “substantive desert thesis,” but I argue that we should not accept this account as a positive characterization of basic desert. However, understanding why is instructive in regards to the distinction between basic and non-basic desert of blame, and also suggests an argument for a new analysis of basic desert of blame. In section 3 I sketch this positive account, the *fittingness account*, and argue that it has two significant explanatory virtues. Finally, in Section 4 I address a familiar problem for analyses that appeal to fittingness in other domains – circularity – and present three options for avoiding this problem as it relates to my proposed account of basic desert of blame.

But isn’t there something strange about trying to explicate basic desert? Doesn’t the very fact that this brand of desert is *basic* preclude this kind of analysis? As McKenna puts the point:

If a concept is basic, then it cannot be derived from other more basic concepts, or it cannot be exhaustively explained (noncircularly) in terms of them, or it cannot be reductively analyzed in terms of them. (2012, p. 121)

True, attempts to analyze a *conceptually* basic primitive would be misguided, but I see no reason to consider the notion of desert at issue to be basic in this sense. In practice, we often

fall back on conceptual primitiveness only after a great deal of work analyzing a concept has failed, and the prospects for future success begin to look reasonably grim. But explicit attempts to analyze basic desert, especially basic desert of blame, are still largely underdeveloped. Perhaps the difficulty in explicating basic desert of blame arises largely from its close relation to a family of other normative concepts including fairness, justification, appropriateness, fittingness, and correctness. At the end of the day it may turn out that basic desert is best understood as *most* basic among these, but even so there might still be a great deal to say about basic desert. Here again McKenna puts the point nicely:

But [desert] can nevertheless be elucidated. What inferences are licensed by basic desert? What entails it, and what does it entail? What is the normative force of the claim that blame is deserved, and what sort of practical punch, so to speak, does deserved blame deliver? If that practical punch involves a harm, what does the harm amount to? What about considerations of proportionality between the wrong serving as the basis of blame and the activity of blaming? (2012, p. 121)

Thus even if basic desert does turn out to be a conceptual primitive, attempting to analyze it might still take us some ways towards *elucidating* basic desert, and towards answering some of the questions above. So, in what follows I will assume that basic desert is not in fact primitive in a sense that would make further analysis misguided.⁹

1 Basic versus Non-basic Desert

Perhaps the best way to get a foothold on what basic desert is begins with what basic desert is not. Most recent attempts to articulate the distinction between basic and non-basic desert take Feinberg's remarks on desert as a starting point:

In respect to modes of treatment which persons can be said to deserve, then we can distinguish three kinds of conditions. There are those whose satisfaction confers eligibility ('eligibility conditions'), those whose satisfaction confers entitlement ('qualification conditions'), and those conditions not specified in any regulatory or procedural rules whose satisfaction confers worthiness or desert ('desert bases'). (Feinberg 1970, p. 58).

Feinberg suggests that the distinction between different modes of desert depends in some way on the conditions whose satisfaction confer desert. According to Feinberg, basic desert is conferred only by the satisfaction of conditions "not specified in any regulatory or procedural rules." In other words, in order for the satisfaction of some conditions to confer basic, rather than non-basic desert, they must be pre-institutional or non-instrumental. Here is how Pereboom puts the point in regards to basic desert of blame in particular:

The desert at issue here is basic in the sense that the agent, to be morally responsible, would deserve to be the recipient of the expression of such an attitude just because she has performed the action, given sensitivity to its moral status; not, for example, by virtue of consequentialist or contractualist considerations. (Pereboom 2012, p. 189).

Feinberg and Pereboom's remarks emphasize the importance of the relevant notion of a *desert base* in articulating the difference between basic and non-basic desert. In what follows I take a desert base to be whatever it is that agents are deserving in virtue of. Feinberg and Pereboom's attempts to distinguish basic from non-basic desert suggest that both kinds of desert have the

same conceptual structure, and that the difference between the two depends on the nature of the relevant desert base. As a first pass then we can understand the conceptual structure of desert as a three-place relation involving an agent (A), an object – what it is that is deserved (X), and a base – what it is that X is deserved in virtue of (B):

- (1) A deserves X in virtue of B.¹⁰

What constitutes the difference between basic and non-basic desert is the nature of the base, B. And initially we can at least specify what the relevant bases *cannot* be like if they are to generate basic, rather non-basic, desert. Again borrowing from Feinberg and Pereboom, institutionally generated, instrumental, contractualist, and consequentialist bases are out. At best they generate only non-basic desert.

This characterization of basic desert is still largely negative, but can be further explicated by restricting the scope of the object of desert at issue. For my current purposes the relevant object is blame. First, shifting attention to basic desert of blame highlights a shortcoming of characterizing the structure of basic desert in terms of (1) above. Understanding basic desert of blame in particular as merely a three place relation is problematic because it fails to recognize a distinction between the subject of desert (in this case an agent) and what it is that subject deserves blame *for*. When we ask whether an agent deserves blame we do not, for example, ask whether she deserves blame in the same generalized way that we might ask whether persons deserve respect. Rather, we want to know whether the agent in question deserves blame *for a particular action, quality of will, or character trait*. Structurally, then, when restricted to blame basic desert is better understood as a four-place relation:

- (2) A deserves blame *for* Y in virtue of B.

Given the relatively uncontroversial features of the distinction between basic and non-basic desert identified above, we should also exclude post-institutional, contractualist, consequentialist, and instrumental considerations as possible bases for basic desert of blame. But once the scope of basic desert is restricted to blame in particular there is more to say about what the relevant desert bases *are*. In light of the fact that basic desert of blame has the structure identified in (2), it is natural to suppose that the kinds of things an agent can deserve blame in virtue of must be features of what it is she deserves blame *for*. The most plausible candidates for desert bases for basic desert of blame, then, will be features of the agent's action, quality of will, or character traits.¹¹

Though rarely made explicit, I take this characterization of basic desert of blame to be accepted by the majority of those interested in basic desert in the context of free will and moral responsibility.¹² It should therefore come as no great surprise that it is also relatively uninteresting. This characterization still leaves open, for example, questions about *which* features of an agent's action, quality of will, or character are candidate desert bases, and *why* these features, but not others, constitute bases for basic desert of blame.

It is not clear how one might answer the former question without presupposing a substantive positive view of the conditions for blameworthiness and moral responsibility, presuppositions I wish to avoid here. However, there is room to explore a version of the second question independent of such presuppositions. Though admittedly natural to suppose, are there any positive arguments in favor of taking only features of an agent's action, quality of will, or character to be the kinds of things agents can deserve blame in the basic sense in virtue of?

Examining where one positive attempt to explicate desert of blame goes wrong suggests that there are.

2 Substantive Desert Theses and the Wrong Kind of Reasons

In this section I take Michael McKenna's (2012) proposal for a substantive desert thesis as a case study. Though I will argue that McKenna's characterization does not provide an adequate account of desert of blame in the basic sense, understanding why is instructive.¹³ In particular, it suggests a positive argument for thinking that the appropriate bases for basic desert of blame should be restricted only to facts about the agent being blamed and her action.

McKenna offers the following substantive desert thesis for blame (for simplicity, call it SDT):

It is a noninstrumental good that, as a response to the meaning expressed in an agent's blameworthy act, that agent experiences the harms of others communicating in their altered patterns of interpersonal relations their moral demands, expectations, and disapproval. Because this is a noninstrumental good, it is permissible to blame one who is blameworthy (2012, p. 141).

McKenna does not initially commit himself to SDT, but rather proposes it as a desert thesis consistent with his conversational theory of moral responsibility and blame. Here it is beyond my current purposes to dive too deeply into McKenna's rich conversational view, but ultimately he tentatively accepts SDT as the best account of the warrant of blame and identifies three reasons for thinking that the harms of blame according to the conversational theory are in fact non-instrumentally good. First, on McKenna's view the harms of blame involve "adversely affecting a distinctive class of welfare interests," those which concern "unfettered exercising of capacities to engage in social intercourse and maintain friendships, freedom from interference in one's personal life, and a sustained level of emotional stability" (2012, p. 167). McKenna argues that these harms are therefore non-instrumentally good because (1) on the conversational model of blame sensitivity to them requires or reveals that the agent being blamed is committed to membership in the moral community, (2) subjecting a wrongdoer to these harms reveals a commitment to morality on the part of the person doing the blaming, and (3) the conversational process that these harms play an integral part in is itself non-instrumentally valuable (2012, pp. 167–171). On McKenna's view it is therefore permissible for members of a shared moral community to blame one another because it is good to do so, and it is good to do so because our blaming practices are a constitutive feature of a conversational activity that is itself non-instrumentally valuable.

McKenna himself acknowledges that one might deny that the goods at issue here are in fact non-instrumentally valuable (170). This is an interesting question and one that deserves further attention, but here I will grant that they are. More interesting to the topic at hand is McKenna's characterization of these considerations as providing as plausible account of the *warrant* for blame. In light of this characterization, how should we understand SDT? Can SDT provide a robust, positive explication of basic desert of blame? Or is it instead a thesis about what *justifies* blame, either in regards to particular attributions or our blaming practices more generally?

The answer seems clearly to be the latter. In order to see why, consider SDT in light of the structure of basic desert of blame identified by (2) above. What, in particular, does McKenna have in mind as candidate desert bases? SDT suggests the following: an agent deserves blame for a particular action or feature of her character in virtue of *the non-instrumental value of the conversational activity constituted by our blaming practices*. But then, given the uncontroversial features of basic desert identified in section 1, SDT is at best a non-basic desert thesis. Even granting that the value of the conversational activity McKenna is interested in is non-instrumental, the desert base identified in SDT is itself consequentialist. According to SDT, agents deserve blame because of the value of this conversational activity (non-instrumental or otherwise), and *not* in virtue of the features of their action or character that they are being blamed *for*.

The key point to emphasize here is that while McKenna's desert thesis might be plausible as an account of the warrant or *justification* for blame,¹⁴ it is not a plausible account of *desert* of blame in the basic sense. And identifying this distinction is vastly helpful in teasing apart the conceptual space between the two closely related notions of justification and desert. SDT grounds what agents deserve blame in virtue of in facts about the value of blaming them – namely that doing so is a constitutive feature of non-instrumentally valuable conversational activity. But this is clearly *the wrong kind of reason*.

Bearing this point out requires a brief aside regarding the nature of blame itself. While there is widespread disagreement about the nature of blame it is relatively uncontroversial to assume that blame either is or is conceptually connected to an *affective attitude*. This assumption fits most naturally with views that take blame to essentially involve the reactive attitudes,¹⁵ but even those accounts of blame that do not take the reactive attitudes as central often acknowledge this affective, attitudinal component as at least partly constitutive of blame.¹⁶ And Given the assumption that resentment and moral anger are at least partially constitutive of blame, there is a clear analogy to draw between the *reasons* for having these attitudes and the reasons for having other affectively colored attitudes. And a great deal of work has already been done articulating the kinds of reasons that are appropriate or *fitting* for having such attitudes.¹⁷ Consider the classic example of admiration. One might have any number of reasons for admiring a particular object, but a distinction can be made between the kinds of reasons that might make admiration *fitting* or appropriate and those that do not. This distinction has been widely characterized in terms of the right versus wrong kinds of reasons for a given attitude, and it is often highlighted via the use of examples of the wrong kinds of reasons. Here is a familiar example: consider a world in which an evil demon has set things up such that those who do not admire the demon will suffer some sort of hideous punishment, though the demon himself possesses no admirable qualities. In such a world you obviously have a reason to admire the demon – failing to do so will land you with a hideous punishment. But this reason is *of the wrong kind*. While the fact that you have a reason to admire the demon in this case might, all else being equal, *justify* your admiration, there is a clear sense in which your admiration is not fitting or appropriate.

Here I wish to suggest another way of putting the point: the demon does not *deserve* your admiration. And we can extend this line of reasoning to particular instances of blame understood in terms of SDT. If McKenna is right and blame is constitutive of a non-instrumentally valuable conversational activity, then for each particular instance of blame there is a *pro tanto* reason to take up the attitudes of moral anger and resentment. Furthermore, in the absence of further reasons that outweigh it, this reason might also plausibly *justify* a particular instance of blame. But if this reason is the *only* one in favor of taking up these attitudes in a particular case then there is still a clear sense in which moral anger and resentment are *not fitting* or, in other words, *not deserved*.

In short, if we understand desert of blame in terms of SDT then particular attributions of blame look to be subject to a wrong kind of reasons problem. Given the assumption that blame is, at least in part, constituted by the affective attitudes of moral anger and resentment it is reasonable to think that, like other such attitudes, blame can be justified by both the right and wrong kind of reasons. But in cases in which the reasons to blame are of the wrong kind (as in instances of blame understood in terms of SDT) there is no reason to think that particular instances of blame, though justified, are *deserved*. And, given the relatively uncontroversial negative features of basic desert of blame articulated in the previous section there is a clear explanation why. According to SDT the desert base for instances of blame will always be consequentialist.

If this is correct, then a closer look at why SDT falls short of a substantive basic desert thesis suggests a potentially fruitful positive characterization of basic desert, a fittingness account:

- (3) A deserves X for Y in virtue of B, where B is only *the right kind of reason*.

But then what do the *right* kinds of reasons look like? Here one might object that a fittingness account of basic desert of blame is unhelpful unless the notion of fit can provide some further traction in identifying what kinds of desert bases do make blame genuinely fitting or deserved. However, even if we cannot explicitly identify the right kinds of reasons for blame, drawing the analogy between blame and other affective attitudes is still vastly helpful in articulating the distinction between basic and non-basic desert of blame. Thus at the very least it provides further content to the largely negative character of basic desert discussed in the previous section. However, understanding basic desert as fit *can* take us a step further in articulating a positive account of basic desert. In particular, a fittingness account of basic desert of blame has at least two powerful explanatory virtues. First, it suggests a positive argument for taking only features of the agent being blamed and her action as appropriate bases for basic desert. Second, it is particularly well suited to explain the conceptual distinction between basic desert of blame and the closely related normative concepts of justification and fairness.

3 Basic Desert as Fittingness

Section 1 identified one of the only widely agreed upon assumptions about what the appropriate bases for basic desert of blame are: features of the agent and her action. In this section I argue that the fittingness account of basic desert of blame has the unique potential to explain *why* the appropriate bases for basic desert of blame should be restricted to these features. This argument appeals to an analogy between blame and the right kinds of reasons for other affective attitudes. Furthermore, the fittingness account has the additional explanatory virtue of allowing us to distinguish between basic desert of blame, justification, and fairness.

Taken together, these arguments provide good reason to take the fittingness account of basic desert of blame seriously.

First, while the distinction between the wrong versus right kinds of reasons for holding a particular affective attitude seems intuitively clear,¹⁸ it is notoriously difficult to articulate what the *right* kind of reasons are for any particular attitude. Consider again the example of admiration. Like examples of non-basic desert, examples of the wrong kind reason for admiring an object are cheap. If the fact that the demon will subject you to some hideous punishment if you fail to admire it is the wrong kind of reason to admire it (it may render your admiration justified, but it cannot render it fitting or deserved), what would the right kind of reason look like? The obvious answer is that *genuinely admirable properties* possessed by the demon would provide the right kind of reason to admire it.¹⁹

One way to cash out this distinction further is to appeal to *content-based* versus *attitude-based* reasons (or, in Parfit's terms, "object-given" and "state-given" reasons).²⁰ Pamela Hieronymi (2005) provides a particularly clear statement of the distinction: object-given or content-related reasons are provided by or have something to do with the object or content of the attitude, while state-given or attitude-related reasons are provided by or have something to do with the state or attitude itself (Hieronymi 2005, p. 441).²¹

Returning to the admiration example, the fact that you will be subject to a hideous form of punishment if you fail to admire the demon provides an attitude-based or state-given reason

to admire it. This reason might ultimately (again, all else being equal) justify your admiration of the demon, but it is irrelevant to whether the cognitive feature of your attitude (something like the judgment, “The demon is admirable”) is correct. On the other hand, a content-based reason to admire the demon is precisely the kind of consideration that *does* (or at least could) count in favor of thinking that this cognitive feature of your attitude is correct. The fact that the demon does in fact possess some admirable feature would provide a content-based reason for you to admire it. So, for those who accept this distinction only content-based reasons provide the right kind of reasons for a particular attitude.

Returning to blame, the distinction between content-based and attitude-based reasons suggests a line of argument for thinking that only facts about the features of an agent and her action constitute the right kind of reasons to blame. Like admiration, blame plausibly has a cognitive component, something along the lines of the judgment, “A is *blameworthy*.”²² And so only considerations that count in favor of the correctness of this judgment will count as content-based reasons to blame. Which considerations are those? In turning to work on blameworthiness, one widely accepted necessary condition for an agent being blameworthy is that she is morally responsible for her action. The jury is, of course, still out on which particular considerations count in favor of thinking an agent is morally responsible for her action. But I take it that all prominent accounts of moral responsibility accept that it will be *some* features of the agent and her action. And so, on a fittingness account of blame, only facts about the features of an agent and her action constitute bases for basic desert of blame, because only these facts provide the right kind of reasons to adopt the relevant attitude. If this is correct, the fittingness account has the theoretical virtue of explaining *why* only features of the agent being blamed and her action constitute bases for basic desert of blame.

Furthermore, the fittingness account has the additional virtue of allowing us to explain the distinction between desert and the closely related concepts of justification and fairness.

These concepts are often run together in matters related to blame and moral responsibility. In regards to justification, it is sometimes assumed (recall the examples of charges often lodged against revisionists and compatibilists discussed at the outset of this paper) that the fact that blame is not deserved in the basic sense entails that our blaming practices or particular instances of blame cannot be justified. But this claim is rarely argued for, and the discussion of SDT in the previous section suggests that there is good reason to think that these concepts come apart. The fittingness account of basic desert of blame not only respects this conceptual distinction, but also provides an explanation for why it exists – basic desert of blame and the justification of blame are generated by *different kinds of reasons*. Perhaps what we are interested in or what we want desert to track for us is something more closely akin to *correctness*, not justification.²³ And thus restricting the bases for basic desert of blame to just those considerations that count in favor of correctness, as the fittingness account recommends, captures this intuition nicely.

What of fairness? It is less obvious that desert and fairness are normatively and conceptually distinct. However, Dana Nelkin has recently argued (I think persuasively) that desert is only one of multiple considerations that bears on the overall fairness of our treatment of one another (2013, p. 122). Furthermore, there are at least some considerations which plausibly undermine desert but not fairness. As Nelkin points out, if an agent has no control over her action this intuitively disqualifies her from genuine desert of sanctions or adverse treatment. But surely there is room to say that it could still be *fair* to sanction her if, for example, the procedure for doing so is fair in other ways (2013, p. 122). If we take blame to be a kind of sanction, then the fittingness account also suggests a potentially fruitful explanation for this

distinction. Basic desert of blame, on the fittingness account, is fairly narrow. The fact that an agent had no control over her action looks like a plausible candidate for a consideration that counts in favor of thinking that the judgment “A is blameworthy” is false. Whatever the necessary conditions for blameworthiness are, some degree of control over one’s action is likely one of them. And so the fittingness account of blame provides an explanation for why the agent who lacks control entirely does not deserve blame in the basic sense. But, importantly, the fittingness account also leaves room to say that, even so, it might be fair to blame her. If, for example, some version of eliminativism about moral responsibility is true, then perhaps all of us lack the kind of control necessary for moral responsibility all of the time. It might nonetheless be fair to continue to go on blaming and sanctioning one another given that we are all subject to the same lack of control. None of this is to say that it would be fair to blame an agent who completely lacks control of her action, but it is a virtue of the fittingness view that it allows us to explain cases in which desert and fairness seem intuitively to come apart. Again, it is because they depend on different kinds of reasons.

4 Circularity

Taken together, the arguments in the previous section provide good reason to take the fittingness account seriously. However, there is a particularly troubling potential objection looming. Fittingness accounts in other domains often face charges of vicious circularity – the concept being analyzed in terms of fit appears somewhere in the explication of the fittingness relation itself. This claim is often lodged against fittingness (or “buck-passing”) analyses of value. Is the fittingness account of basic desert of blame open to this kind of objection?

The severity of circularity objections to fittingness accounts often depends on the level of generality of the thing being analyzed. The more general the concept being analyzed in terms of fit, the more likely it is that this very concept will appear somewhere in the explication of the fittingness relation. Fittingness accounts of value, for example, look to be especially vulnerable to circularity worries. If value is to be analyzed in terms of pro-attitudes that are fitting, what makes certain pro-attitudes, but not others, fitting? Appeal to the fact that the fitting attitudes are those generated by content-based reasons, or considerations that count in favor of thinking the judgment “X is valuable” are correct, look viciously circular. We are left analyzing value in terms of a particular set of fitting pro-attitudes, but the fact that these pro-attitudes are fitting is grounded in the fact that the object of these attitudes is valuable. Is the fittingness account of basic desert of blame open to the same kind of objection?

Circularity worries about fittingness analyses often dissipate as the specificity of the thing being analyzed increases, and the fact that the fittingness account on offer is relatively specific – it is an attempt to explicate not only basic desert, but basic desert of *blame* in particular – itself undermines the force of circularity worries here. However, in light of its force in other domains this worry still constitutes a serious obstacle to motivating the fittingness account of basic desert of blame, and so in what remains I argue that there is good reason to think it can be addressed.

Applied to the fittingness account of basic desert of blame, the circularity objection has the following structure. The fittingness account analyzes basic desert of blame in terms of the right kind of reasons to blame. But what makes a reason to blame of the right kind is just the fact that the agent being blamed is blameworthy. So the fittingness account recommends that we analyze blame in terms of blameworthiness, which is at best uninformative and at worst viciously circular.

This initial formulation of the objection rests on a confusion about what is being analyzed. In this case it is not blame itself, but what blame is *deserved in the basic sense in virtue of* that is being analyzed in terms of fit. The circularity objection in its basic form would only be relevant to the fittingness account on offer if *basic desert* were being analyzed in terms of the right kind of reasons, and the right kind of reasons analyzed in terms of, for example, what is genuinely *deserved*. But that is not the proposal. The fittingness account does take basic desert to be analyzed in terms of the right kind of reasons, but on this account the right kind of reasons are analyzed in terms of blameworthiness. And so this version of the circularity objection is not relevant to the fittingness account of basic desert of blame.

However, there is a version of the objection that is relevant. The proponent of this objection can and should press the point that the above response conveniently stops a step short in the analysis. How are we to analyze blameworthiness? For the fittingness account of blame to ultimately be informative and avoid the circularity worry, it must be possible to provide an account of blameworthiness that does not itself appeal to desert. Given that the examples used to motivate many accounts of blameworthiness *do* often appeal to intuitions about whether or not agents are deserving of blame, is it plausible to think that a successful account of blameworthiness can avoid such appeals?

Responding to this objection fully would require a full blown positive account of blameworthiness, and so here I can only attempt to undermine its force by identifying at least three options available to the defender of the fittingness account. The first option is to provide an account of blameworthiness that is largely *descriptive*. This option is Strawsonian in spirit. Here the proponent of the fittingness account might assume that it is possible to provide a successful analysis of blameworthiness based entirely on our blaming practices as we find them. It is not obvious that such an analyses need make any appeal to desert of blame. Second, a defender of the fittingness account might assume that it is possible to provide a successful analysis of blameworthiness that is largely *realist*. Perhaps metaphysical facts about blameworthiness are simply part of the fabric of the world, independent of creatures like us and our intuitions about desert. Finally, one might defend the fittingness account via appeal to the possibility of providing a successful analysis of blame that is largely *normative*, but independent of facts about desert. Perhaps such an analysis can succeed via appeal only to other normative concepts. One obvious candidate concept is *fairness*. Given that the fittingness account is particularly well suited to treat fairness as a concept independent of desert I take this third option to be especially fruitful.

5 Conclusion

We now have in hand a proposal for what the shape of the elephant in the room looks like – basic desert of blame can be understood in terms of the fittingness of its constitutive attitudes of moral anger and resentment, the fittingness of these attitudes (or the bases appropriate to basic desert of blame) can be analyzed in terms of the right kind of reasons, and there are at least three viable options for analyzing the relevant reasons in terms of blameworthiness that are not viciously circular or uninformative. At the very least, the fittingness account helps to provide further content to characterizations of basic desert that have thus far been largely negative. At best, it is a novel and potentially fruitful positive analysis of basic desert of blame, one that has the powerful theoretical virtues of explaining not only why the only bases appropriate to basic desert of blame are features of the agent being blamed and her action, but also the distinction between desert and other closely related normative concepts like justification and fairness.

Notes

- 1 For a notable exception, see Haji's arguments against (BRI) in this volume.
- 2 By *success theory* I mean views that accept necessary and sufficient conditions for moral responsibility which, taken together, human beings do sometimes satisfy in the actual world. Thus on a success theory we are, at least sometimes, morally responsible for our actions.
- 3 Here I have in mind views that accept necessary and sufficient conditions for moral responsibility which human beings don't (or can't) ever satisfy in the actual world. See, for example, Derk Pereboom's hard incompatibilism (2001, 2007, 2014), and Galen Strawson's impossibilism (1994).
- 4 For the original presentation of this argument see Pereboom (2001).
- 5 However, Daniel Haas (2013) has recently argued that many Strawsonian compatibilists may not presuppose the same notion of basic desert that Pereboom saddles them with, and further that it is an open question whether or not they should. Haas also argues for a distinction between basic desert understood as fit and basic desert understood as merit. According to Haas's interpretation, the kind of basic desert Pereboom stipulates is at issue in the debate between compatibilists and incompatibilists is merit-based desert. But, Haas argues, there is also a robust notion of fit-based desert which, although perhaps not basic in the non-comparative, pre-institutional sense, might still be a robust notion of basic desert. I discuss these distinctions further in section 1, and here wish to point out that the fittingness account defended in what follows is intended to be the kind of desert Pereboom stipulates is at issue in these debates.
- 6 See Pereboom (2001, 2005, 2007, 2009a, 2009b, 2014), and Vargas (2005, 2009, 2011, 2013).
- 7 See Pereboom (2009b) and McKenna (2009).
- 8 Thanks to Michael McKenna for suggesting the latter alternative.
- 9 Though perhaps at the end of the day the lesson to draw from attempts like mine will be that this assumption is false. One reason to think basic desert *is* primitive in a sense that would make such explication misguided might be intuitive worries about circularity. I address this kind of worry in section 4 and argue that there are at least three plausible strategies for avoiding it on the fittingness account of basic desert of blame I offer.
- 10 Here I borrow from Roskies and Malle's (2013) helpful articulation of what they call the "conceptual structure" of basic desert.
- 11 In Section 3 I provide an argument for thinking that these features *should* be taken as the only things agents can deserve blame in the basic sense in virtue of. In this section my goal is only to canvas the few widely agreed upon features those interested in basic desert often *do* appeal to in order to distinguish basic from non-basic desert.
- 12 See, for example, Pereboom (2001, 2009b, 2012). This notion of basic desert also appears to be what Galen Strawson has in mind when he talks about agents being "truly and without qualification deserving of praise and blame or reward or punishment" (2002, p. 442). McKenna also seems to accept this characterization of basic desert of blame, though he claims that *as specified* it is "too lean to assess" (2012, p. 126). I agree that *as specified* it is too lean, but think that there is more to be said about genuinely basic desert and develop this view in what follows.
- 13 McKenna himself does not explicitly call this a *basic* desert thesis. Here I am not interested in offering an objection to McKenna's characterization, only in whether or not his desert thesis *should* be understood as a basic desert thesis, and argue that it should not be.
- 14 Or perhaps even a kind of forensic evidence in support of the claim that there is some non-instrumental good in harming a person by blaming her. Thanks to McKenna for suggesting this further possibility.
- 15 See Strawson (1962) and Wallace (1994, 2011). Variants include Cohen (1977), Fingarette (1957), Wertheimer (1998), and Wolf (2011).
- 16 See for example McKenna's (2012, 2013) communicative view, and Franklin's (2013) functionalist view. Smith (2013) and McGeer (2013) also offer functionalist accounts that take the reactive attitudes to play a central role, though they each hold that these attitudes are not a necessary feature of every instance of blame. T.M. Scanlon's (2008) view, as well as George Sher's (2006),

are two notable examples of views that do not that the reactive attitudes to feature in the nature of blame. However, such views are often charged with the objection that in leaving out the reactive attitudes they fail to capture the “force” of blame (Wolf 1990; Hieronymi 2004), or are too “sanitized” (McGeer 2013).

- 17 See, for example, Brentano (1969), D’Arms and Jacobsen (2000a, 2000b), Olson (2004), Rabinowicz and Rönnow-Rasmussen (2004, 2006), Suikkanen (2004), and Hieronymi (2005).
- 18 Though some do in fact reject this distinction. See Rabinowicz and Rönnow-Rasmussen (2004).
- 19 This initial proposal will likely flag the circularity worries mentioned at the outset of this paper. I address these worries in the following section.
- 20 Here I intend to use this particular distinction largely as a heuristic. I suspect that a satisfying application of the distinction in this particular context requires a much more fine-grained account of the attitudes of moral anger and resentment, an account which lies well beyond the scope of this paper. However, for those who reject this distinction, I think that the same line of argument can be run in terms of epistemic and pragmatic reasons (though see Reisner (2009) for dissenting arguments). For those who reject this kind of distinction altogether, the fittingness account on offer will likely be deeply unappealing from the outset.
- 21 I also credit Hieronymi (2004) with offering a precursor to the fittingness account I develop here, and for initially suggesting that this kind of distinction might be applied to blame and in particular resentment.
- 22 For example, see Sher’s (2006) discussion of the belief component of blame. Sher, however, accounts for the force of blame via appeal to a corresponding desire rather than (like the fittingness account on offer) our affectively colored reactive attitudes.
- 23 Hieronymi (2004) also suggests we take seriously the same kind of gap regarding the appropriateness of resentment.

Bibliography

- Brentano, F. (1969) [1889]. *The Origin of Our Knowledge of Right and Wrong* (ed. O. Kraus and R. Chisholm, trans. R. Chisholm and E. Schneewind). London: Routledge & Kegan Paul.
- Cohen, S. (1977). Distinctions among blame concepts. *Philosophy and Phenomenological Research* 38: 149–166.
- D’Arms, J. and Jacobsen, D. (2000a). The moralistic fallacy: on the ‘appropriateness’ of emotions. *Philosophy and Phenomenological Research* 61: 65–90.
- D’Arms, J. and Jacobsen, D. (2000b). Sentiment and value. *Ethics* 110: 722–748.
- Feinberg, J. (1970). *Doing and Deserving*. Princeton, NJ: Princeton University Press.
- Fingarette, H. (1957). Blame: its motive and meaning in everyday life. *Psychoanalytic Review* 44: 193–211.
- Franklin, C. (2013). Valuing blame. In: *Blame: Its Nature and Norms* (ed. J. Coates and N. Tognazzini), 207–223. Oxford: Oxford University Press.
- Haas, D. (2013). Merit, fit, and basic desert. *Philosophical Explorations* 16 (2): 226–239.
- Hieronymi, P. (2004). The force and fairness of blame. *Philosophical Perspectives* 18: 115–148.
- Hieronymi, P. (2005). The wrong kind of reason. *The Journal of Philosophy* 102: 437–457.
- McGeer, V. (2013). Civilizing blame. In: *Blame: Its Nature and Norms* (ed. J. Coates and N. Tognazzini), 162–188. Oxford: Oxford University Press.
- McKenna, M. (2009). Compatibilism and desert: critical comments on *Four Views on Free Will*. *Philosophical Studies* 144: 3–13.
- McKenna, M. (2012). *Conversation and Responsibility*. Oxford: Oxford University Press.
- McKenna, M. (2013). Directed blame and conversation. In: *Blame: Its Nature and Norms* (ed. J. Coates and N. Tognazzini), 119–140. Oxford: Oxford University Press.
- Nelkin, D. (2013). Desert, fairness, and resentment. *Philosophical Explorations* 16 (2): 1–16.
- Olson, J. (2004). Buck-passing and the wrong kind of reasons. *The Philosophical Quarterly* 54: 295–300.
- Pereboom, D. (2001). *Living without Free Will*. Cambridge: Cambridge University Press.

- Pereboom, D. (2005). Defending hard incompatibilism. *Midwest Studies in Philosophy* 29: 228–247.
- Pereboom, D. (2007). Hard incompatibilism, and response to Kane, Fischer, and Vargas. In: *Four Views on Free Will* (ed. R. Kane, J.M. Fischer, D. Pereboom and M. Vargas), 85–125 and 191–203. Oxford, UK: Blackwell.
- Pereboom, D. (2009a). Free will, love, and anger. *Ideas y Valores* 141: 5–25.
- Pereboom, D. (2009b). Hard incompatibilism and its rivals. *Philosophical Studies* 144: 21–33.
- Pereboom, D. (2012). Free will skepticism, blame, and obligation. In: *Blame: Its Nature and Norms* (ed. J. Coates and N. Tognazzini), 198–206. Oxford: Oxford University Press.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
- Rabinowicz, W. and Rønnow-Rasmussen, T. (2004). The strike of the demon: on fitting pro attitudes and value. *Ethics* 114: 391–423.
- Rabinowicz, W. and Rønnow-Rasmussen, T. (2006). Buck-passing and the right kind of reasons. *The Philosophical Quarterly* 56: 114–120.
- Reisner, A. (2009). The possibility of pragmatic reasons for belief and the wrong kind of reasons problem. *Philosophical Studies* 145: 257–272.
- Roskies, A. and Malle, B. (2013). A Strawsonian look at desert. *Philosophical Explorations* 16 (2): 133–152.
- Scanlon, T.M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Cambridge University Press.
- Sher, G. (2006). *In Praise of Blame*. Oxford: Oxford University Press.
- Smith, A. (2013). Moral blame and moral protest. In: *Blame: Its Nature and Norms* (ed. J. Coates and N. Tognazzini), 27–48. Oxford: Oxford University Press.
- Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies* 75: 5–24.
- Strawson, G. (2002). The bounds of freedom. In: *The Oxford Handbook of Free Will* (ed. R. Kane), 441–460. New York: Oxford University Press.
- Strawson, P.F. (1962). Freedom and resentment. *Proceedings of the British Academy* 48: 1–25.
- Suikkanen, J. (2004). Reasons and value: in defense of the buck-passing account. *Ethical Theory and Moral Practice* 7: 513–535.
- Vargas, M. (2005). The revisionist's guide to responsibility. *Philosophical Studies* 125 (3): 399–429.
- Vargas, M. (2009). Revisionism about free will: a statement and defense. *Philosophical Studies* 144 (1): 45–62.
- Vargas, M. (2011). Revisionist accounts of free will: origins, varieties, and challenges. In: *The Oxford Handbook of Free Will 2nd Edition* (ed. R. Kane), 457–474. Oxford: Oxford University Press.
- Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.
- Wallace, R.J. (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.
- Wallace, R.J. (2011). Dispassionate opprobrium: on blame and the reactive sentiments. In: *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon* (ed. R.J. Wallace, R. Kumar and S. Freeman), 348–372. New York: Oxford University Press.
- Wertheimer, R. (1998). Constraining condemning. *Ethics* 108: 489–501.
- Wolf, S. (1990). *Freedom within Reason*. New York: Oxford University Press.
- Wolf, S. (2011). Blame, Italian style. In: *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon* (ed. R.J. Wallace, R. Kumar and S. Freeman), 332–347. New York: Oxford University Press.

Criminal Responsibility

KEN LEVY

1 Introduction

This entry will explicate the conditions required for criminal responsibility, provide an overview of criminal defenses, distinguish criminal responsibility from both tort liability and moral responsibility, and explicate the current state of the insanity defense.

2 Actus Reus, Mens Rea, Strict Liability, and Causation

In American criminal law, an individual is considered criminally responsible – responsible for committing a crime – when a factfinder (judge or jury) determines that the elements of a criminal statute are all satisfied.

One element common to every criminal statute is *actus reus*, the prohibited action or omission. The action or omission must be voluntary, which is generally taken to mean that the individual herself – as opposed to a muscle twitch or spasm – “caused” the bodily motion that constituted the action or omission. We normally refer to this kind of causation as *control* or *agency*. Agency implies that the individual was not forced to act or omit as she did and therefore that she could have done otherwise – that is, she could have refrained from performing the action or omission. Importantly, while possession is not generally considered to be an action per se, it can still qualify for *actus reus* because it requires a previous action (acquiring) or previous omission (failing to relinquish).

In most criminal statutes, a second element – the *mens rea*, a certain state of mind that accompanies the prohibited action – must also be satisfied. There are four kinds of *mens rea* (*mentes rea*): specific intent, knowledge, recklessness, and negligence. I say *most*, not all, criminal statutes require *mens rea* because a small number of criminal statutes are a matter of *strict liability*; they do not require intent, knowledge, recklessness, or even negligence. Instead, they require only *actus reus*. For example, many rape statutes impose liability on

adults for engaging in sexual relations with minors even if the former's mistaken belief that the latter has reached the age of majority is reasonable. Still, most strict liability crimes are insignificant – that is, either misdemeanors or regulatory (public welfare) offenses for which relatively small fines are imposed. Two examples are speeding and selling alcohol to a minor.

In most criminal statutes, the required mens rea is *specific intent*, which is equivalent to conscious purpose. In many criminal statutes, the mens rea is a state of mind that is easier to prove than specific intent: knowledge or recklessness. Knowledge is awareness of a particular fact – either of the act itself, of a circumstance occasioning the act, or of a practically certain consequence of the act. Recklessness is also a kind of knowledge – not knowledge of a particular *fact* but rather knowledge of a substantial and unjustifiable *risk* of causing serious harm.

In a small number of criminal statutes, only negligence is required. In most of these statutes, negligence means either (a) inadvertent (as opposed to conscious or knowing) disregard of a substantial and unjustifiable risk of causing serious harm or (b) gross disregard of the need to use reasonable care in order to avoid causing serious harm. (a) is considered to be “recklessness lite” – that is, recklessness minus (evidence of) contemporaneous awareness of the risk; (b) is considered to be a conscious or unconscious deviation from the ordinary standard of care that is *gross* – that is, significant in degree. (a) and (b) are usually taken to be identical, either in meaning or scope.

In some jurisdictions, courts have decided that *civil* negligence is sufficient for such crimes as pollution and vehicular manslaughter. Civil negligence is the standard of liability in tort law; it involves only a deviation – as opposed to a *gross* deviation – from the ordinary standard of care. To the extent that civil negligence has bled into criminal law, the boundary between tort law and criminal law has blurred.

Finally, many criminal statutes require a third element: satisfaction of a particular circumstance. For example, all homicide statutes contain a causation element. In order to be guilty of murder, manslaughter, or negligent homicide, a defendant must satisfy not merely the *actus reus* (for example, shooting the victim) and the *mens rea* (for example, intending to kill the victim) but also causation; the bullet fired from the defendant's gun must cause the victim's death. In order for the shooting to qualify as the cause of the victim's death, it must satisfy both an empirical element (the “factual” cause) and a normative element (the “legal” or “proximate” cause). In order to qualify as a factual cause, the shooting must have been either a necessary condition of the victim's death or part of a group of causes that itself was a necessary condition of the victim's death. In order to qualify as a proximate cause, the shooting must have been closely connected to the victim's death. For example, the shooting would probably *not* qualify as the proximate cause of the victim's death if it merely traumatized her and, as a result of the trauma, she commit suicide two years later. Still, unlike the factual cause, this determination would be a matter of subjective or intersubjective judgment, not a matter of objective fact (such as the bullet's exact trajectory).

3 Criminal Defenses

Suppose that an individual – Mikey – has been arrested for crime C. At some point, Mikey will talk with a prosecutor (local, state, or federal, depending on the crime charged), and the prosecutor will most likely “plea bargain” with him – that is, try to work out a deal with Mikey in order to avoid a trial, which tends to be very costly for both sides. But if they fail to work

out a deal and the case does proceed to trial, the prosecutor has the burden of proof; she must provide evidence to the factfinder proving beyond a reasonable doubt that the defendant did in fact commit *C*. If the crime is strict liability, the prosecutor will need to establish only the required *actus reus*. Otherwise, the prosecutor will also need to establish the required *mens rea* and every other element in the statute.

Once the prosecutor has provided sufficient evidence to establish the defendant's guilt, the burden of proof shifts to Mikey. He now has four possible responses – defenses – available to him. The first possible defense is that the prosecutor failed to prove at least one of the elements, and therefore his guilt, beyond a reasonable doubt.

The second possible defense is admitting to the act in question but denying that it was a crime on the grounds that the act was justified – that is, right, good, or permissible. The traditionally recognized justifications are consent, necessity, self-defense, and defense of others.

The third possible defense is admitting to the crime but claiming blamelessness because, given the particular circumstances, he cannot be reasonably expected to have avoided committing the crime. The traditionally recognized excuses are automatism, duress, entrapment, infancy, insanity, involuntary intoxication, mistake of fact, and mistake of law.

The fourth possible defense: Instead of claiming full exculpation, Mikey might plead a mitigating factor, a condition that makes the criminal act somewhat understandable and thereby reduces the level of the offense for which he is convicted. Traditionally recognized mitigating factors include addiction, extreme emotional disturbance (EED), mental illness, past abuse or neglect, and provocation (“heat of passion”). If Mikey is convicted, then – depending on the circumstances – he may offer several other mitigating factors prior to sentencing: elderly status, first-time offender, minor role in the crime, no harm caused, no longer dangerous, physical illness, or remorse.

4 Criminal Law vs. Tort Law

Once Mikey either pleads guilty to *C* or is found guilty by a factfinder of committing *C*, the judge will try to determine an appropriate sentence. The two most common forms of criminal punishment are imprisonment and fines. While punishment is generally thought to be the defining feature of criminal law, other areas of law, principally torts and administrative law, also prescribe punishment for certain transgressions. Juries sometimes impose punitive damages on tortfeasors, and public employees who violate administrative laws can be punished in a variety of ways, ranging from censure and suspension at one end to fines and imprisonment at the other.

In tort law, punitive damages are rare. The much more typical remedies are restitution (repayment) and compensatory damages. Neither restitution nor compensatory damages are thought to qualify as *punishment* because their primary purpose is not retributive; they are not designed to inflict suffering, hardship, or deprivation on the tortfeasor *as an end in itself*. Rather, their primary purpose is to make the victim whole, to restore her to the condition that she was in prior to being injured. Of course, restitution and compensatory damages cannot always restore health, no less life or limbs. But they are nonetheless considered to be the next best thing when full qualitative restoration is impossible.

Still, restitution and compensatory damages are like punishment in three respects. First, they are all designed to inflict deprivation on the guilty party. (Again, such infliction is an intrinsic end for punishment, an instrumental end for restitution and compensatory

damages.) Second, they are all designed to achieve “general deterrence” (discouraging the general public from engaging in the same kind of illegal conduct) and “specific deterrence” (discouraging the guilty party himself from recidivating). Third, they are all designed to express disapproval of the guilty party’s offense. The level of disapproval is reflected by the severity of the sentence or assigned damages, which itself is supposed to be roughly proportional – or at least not grossly disproportionate – to the severity of the offense.

5 Criminal Responsibility vs. Moral Responsibility

Criminal responsibility and moral responsibility are clearly distinct things because we can have one without the other. A person is morally responsible but not criminally responsible in two situations: (a) she commits an act that is immoral but not illegal or (b) she commits an illegal act, but the criminal justice system does not hold her accountable for whatever reason – for example, she was never caught, the prosecutor offered her immunity in exchange for her assistance, or the crime was *de minimis*. One might argue that (b) involves *unrecognized* criminal responsibility, but ordinary usage of the term *criminal responsibility* implies a formal finding of guilt.

(a) and (b) seem pretty straightforward, but what about the converse of (a) and (b)? Can a person be criminally responsible without also being morally responsible? It is generally assumed that the answer is no, that criminal responsibility *requires* moral responsibility. But this assumption is false. There are at least three situations in which a person might be criminally responsible without also being morally responsible: erroneous conviction, psychopathy-motivated criminality, and situationism.

Suppose an individual – Norris – leads a very clean life. He works hard at his job, takes good care of his young children, treats everybody with kindness and respect, and does not engage in any illegal activity whatsoever. One day, Norris’s laptop suddenly crashes, so he drops it off at Best Buy for repair. As it turns out, the repair guy at Best Buy – Oliver – has it in for Norris. After repairing Norris’s laptop, he plants some child pornography on the hard drive, returns it to Norris, and then alerts the FBI. The next day, the FBI knock on Norris’s door, he invites them in, they ask to see his computer, Norris consents, they find the child pornography, and they immediately arrest Norris for possession of child pornography, which is prohibited by 18 U.S.C. § 2252A.

Norris is clearly innocent – morally. But he may still be *criminally* responsible if either he pleads guilty (most likely to avoid a trial and therefore the risk of much greater punishment) or a factfinder believes (wrongly) that the prosecution has proven all the elements of § 2252A, including knowledge, beyond a reasonable doubt. Both of these situations exemplify criminal responsibility without moral responsibility and therefore the proposition that criminal responsibility does not require moral responsibility.

In *Free Will, Responsibility, and Crime: An Introduction*, I argue that criminal responsibility and moral responsibility may come apart in two other kinds of situation as well. One is when a clinical psychopath – Psycho – commits a violent crime *C* as a result of his psychopathy. Arguably, Psycho is not morally responsible for *C* because, as a psychopath, he does not really understand that, or why, *C* is morally wrong. But Psycho is still *criminally* responsible for *C* because, despite his moral ignorance, he intentionally *C*-ed knowing full well that *C* was illegal and that, if caught, he would likely be arrested and punished for *C*-ing.

The second case involves situationism, the social-psychological theory that environmental circumstances sometimes play a more significant role in explaining an individual’s behavior

than his character. Return to Oliver, who planted child pornography on Norris's computer. On the surface, it looks like Oliver is just a bad guy – that is, has bad character. But we can imagine a situation in which Oliver is actually a good guy – has good character – but still frames Norris because of external pressures. Suppose, for example, that Oliver is in desperate financial straits, is terrified that he will lose his job and not be able to find another, is instructed by his boss – Quincy – to frame Norris, and very reluctantly complies. Arguably, Oliver is not morally responsible. Given his situation, it is not clear that he could have done otherwise, that he had the strength to defy Quincy. But if the police caught and arrested Oliver for planting child pornography on Norris's laptop, his defense that he was “just following orders” would not work. Employer's instructions themselves do not exonerate; only instructions accompanied by credible threats of violence or serious property damage qualify for the duress excuse. Because Oliver would be criminally responsible without necessarily being morally responsible, it follows, once again, that criminal responsibility does not require moral responsibility.

6 The Insanity Defense

In Part 3, I briefly mentioned the traditionally recognized excuses. Once again, they are automatism, duress, entrapment, infancy, insanity, involuntary intoxication, mistake of fact, and mistake of law. All of these are (a) conditions or circumstances (b) that make it unreasonable to expect the agent to have refrained from committing the crime in question (C) and therefore (c) unfair to blame or punish the agent for C-ing. The reason that these particular conditions or circumstances make it unreasonable to expect the agent to have done otherwise is that they conflict with at least one of three conditions necessary for blameworthiness: knowledge, self-control, and free choice. Mistakes of fact and of law conflict with knowledge, automatism conflicts with self-control, infancy and involuntary intoxication conflict with both knowledge and self-control, and duress and entrapment conflict with free choice.

Insanity conflicts with either knowledge or self-control. One version of the insanity defense, the M'Naghten Rule, is purely cognitive, not volitional. It suggests that a mental illness or disability renders a person blameless for criminal activity when it substantially impairs her ability to know right from wrong.

A second version of the insanity defense, the Model Penal Code (or MPC) Rule, contains both a cognitive prong and a volitional prong. (See Model Penal Code § 4.01 (1962).) It suggests that a mental illness or disability renders a person blameless for criminal activity when it negates either her “substantial capacity ... to appreciate” the difference between right and wrong or her “substantial capacity” to act on this understanding and comply with the law. A few jurisdictions have adopted the (once again) purely cognitive M'Naghten Rule and supplemented it with an MPC-like volitional prong, which is commonly referred to as the “Irresistible Impulse Rule.”

Satisfaction of either the M'Naghten Rule or the cognitive prong of the MPC Rule is relatively straightforward.¹ A mental illness or disability impairs an individual's ability to distinguish right from wrong when it causes her to believe, mistakenly, that her action is right, good, or permissible. There are three kinds of situations in which the afflicted individual suffers from this “normative delusion”: the mental illness or disability causes her to hallucinate, to perceive a danger that is actually non-existent, or to adopt beliefs that radically diverge from commonly accepted moral or empirical assumptions. An example of the first situation: killing the neighbor because voices commanded her to. An example of the second

situation: killing the neighbor from a conviction that he posed a lethal threat to herself, another, or the world.

The third situation is harder to explain because of both the conceptual and epistemic difficulties in distinguishing between “crazy” beliefs and radically unpopular but “non-crazy” beliefs. For example, suppose that a person *P* killed a slaveowner in the antebellum south because of her moral conviction that slavery is wrong. *P*’s moral belief here was crazy in the sense of radically unpopular but not crazy from an objective moral perspective. Likewise, a good number of individuals in the United States today are committed neo-Nazis; they are convinced that nonwhites and Jews are both subhuman and dangerous. While mental illness or disability may cause some to subscribe to this radical ideology, mere subscription to this radical ideology is not necessarily a sign of mental illness or disability. Many of them believe in the inherent inferiority and dangerousness of others not because of mental illness or disability but rather because of indoctrination, “tribal” influences and pressures, or a toxic combination of ignorance, anger, and fear.

So what is the difference between a crazy belief and a radical but non-crazy belief? It is not clear that there is one, at least not an intrinsic difference. Radical belief is sometimes, but not always, a sign of mental illness or disability. When determining whether a defendant is insane, a jury needs to determine not merely what radical beliefs the defendant possesses but also *why* the defendant possesses these radical beliefs. And this causal determination will be determined primarily by expert witnesses who specialize in abnormal psychology. If the cause seems to be mental illness or disability, then the defendant will have a stronger case for insanity. Otherwise, if the cause seems to be something else (for example, indoctrination), then she will have a weaker case for insanity.

Whether a defendant has satisfied the Irresistible Impulse Rule (which, again, supplements some M’Naghten statutes) or the volitional prong of the MPC Rule is harder for juries to determine than whether the cognitive prong has been satisfied because we cannot measure self-control with the same degree of precision that we can measure cognitive capacity. For example, if a defendant pleads that (a) he killed another person because he just couldn’t help himself and (b) he just couldn’t help himself because of either a sick compulsion or a psychotic episode, how is the jury supposed to evaluate this claim? Even if the defendant genuinely believes (a) and (b), his belief may be wrong. He may have been able to refrain from killing either by exerting greater willpower at the time or earlier in the day by, for example, taking his medications or consciously avoiding triggers. Again, how is the jury able to determine whether he could have exerted greater willpower or whether medications or consciously avoiding triggers would have been successful? It seems that the Irresistible Impulse Rule and volitional prong of the MPC Rule require juries to engage in extensive counterfactual speculation. And both the quality and results of this counterfactual speculation are all the more questionable if expert witnesses offer conflicting testimony, which is common. For this reason, some scholars think that the Irresistible Impulse Rule and MPC Rule’s volitional prong should be abandoned and the insanity defense remain purely cognitive.

It should be noted that not every state provides defendants with the opportunity to plead the insanity defense. One of them, Montana, abolished it in 1979. The other three – Idaho, Kansas, and Utah – abolished theirs after John Hinckley, Jr., who tried to assassinate President Reagan in 1981, successfully pled insanity in 1982. All four states felt that the insanity defense gave some of the most violent criminals an easy way out; all they needed to do was feign craziness and fool the jury into acquitting them in order to get back

out “on the streets” and commit more violent crimes. Whatever the merits of this concern, the other forty-six states still provide for the insanity defense because of their adherence to two basic precepts: (a) a defendant is not culpable for his behavior if he is insane at the time; and (b) if a defendant is not culpable for his crime, then it is unjust to blame and punish him for it.

While these two precepts seem incontrovertible, the United States Supreme Court seems to have rejected them in *Kahler v. Kansas*, 1405 cf. (02) (2020). In a 6–3 decision, the majority reasoned that, despite first appearances, Kansas has not really abolished the insanity defense in the first place. Instead, they claimed, Kansas has retained it in two ways. First, it provides for a diminished-capacity defense, which allows the defendant to plead that her mental illness or disability prevented her from forming the mens rea required for the crime charged. Second, Kansas allows the defendant to offer mental-health evidence during the sentencing phase of his trial.

The problem with the majority’s decision is that it effectively invites Kansas and the three other abolitionist states – and any other state that now decides to abolish the insanity defense – to violate precepts (a) and (b) above. As the three dissenters (Justices Breyer, Ginsburg, and Sotomayor) argued, and as I argue in Levy (2020b), the diminished-capacity defense is not an adequate substitute for the insanity defense. The reason is that mens rea, which is what the diminished-capacity defense is all about, is not equivalent to culpability, which is what the insanity defense is all about. While an insane individual is, by definition, not culpable for her behavior, many insane individuals can still form the mens rea required for various crimes. From this false equivalence between the insanity and diminished-capacity defenses follow two bad results. The first is that the defendant who successfully pleads diminished capacity is still being blamed and punished, just to a lesser degree. And blame and punishment to *any* degree is unjust if the defendant was not at all responsible for her crime, which is the case if she was insane. The second is factual inaccuracy. By suggesting that a delusional defendant is entitled to diminished-capacity mitigation for (say) a violent crime, a jury is falsely implying that the defendant could not possibly have intended violence.

7 Conclusion

Criminal responsibility – culpability – is a foundational feature of every criminal justice system. As the United States strives to make its own criminal justice system less punitive, less costly, and more equitable, it should look to other countries for alternative perspectives on this universal concept. These alternative perspectives are not merely theoretical; they also have substantial, tangible ramifications for policing, criminal procedure, sentencing, prison conditions, post-conviction relief, and social attitudes.

Note

- 1 Conceptually straightforward, that is. The standards of proof for the M’Naghten Rule, the Irresistible Impulse Rule, the MPC Rule’s cognitive prong, and the MPC Rule’s volitional prong vary across jurisdictions. Usually it is clear and convincing evidence or preponderance of the evidence.

Bibliography

- Bonnie, R., Jeffries, J.C., Jr., and Low, P.W. (eds.) (2008). *A Case Study in the Insanity Defense: The Trial of John W. Hinckley, Jr.*, 3e. New York: Foundation Press.
- Chiesa, L.E. (2014). *Substantive Criminal Law: Cases, Comments and Comparative Materials*, 1e. Durham, NC: Carolina Academic Press.
- Elliott, C. (1996). *The Rules of Insanity: Moral Responsibility and the Mentally Ill Offender*. Albany, NY: State University of New York Press.
- Hafemeister, T.L. (2019). *Criminal Trials and Mental Disorders*. New York: NYU Press.
- Levy, K.M. (2020a). *Free Will, Responsibility, and Crime: An Introduction*. New York: Routledge.
- Levy, K.M. (2020b) Normative Ignorance: A Critical Connection Between the Insanity and Mistake of Law Defenses, *Florida State University Law Review*, 47, 411–443.
- Levy, K.M. and Cohen, A. (2016). Mental Illness, Dangerousness, and Involuntary Civil Commitment. In: *Philosophy and Psychiatry: Problems, Intersections and New Perspectives* (ed. G. Gala and D.D. Moseley), 147–160. New York: Routledge.
- Sinnott-Armstrong, W. and Levy, K. (2011). Insanity Defenses. In: *Oxford Handbook on the Philosophy of the Criminal Law* (ed. J. Deigh and D. Dolinko), 299–334. New York: Oxford University Press.

Part V

The Future

The Experience of Free Agency

OISÍN DEERY AND EDDY NAHMIAS

1 Introduction

Imagine the following situation:

It's your day off. Your phone buzzes. "Oh no," you think, "It's Liam, at work." You're about to head out to play soccer with your kids. You had worried all morning that Liam might not be ready for that big client meeting today and would need your help, even though you had prepared all the relevant files ahead of time. "Surely nothing can go wrong," you think. Yet, when Liam last called on your day off, you had to go in. You've been promising the kids for days you'd play soccer with them. They're waiting outside. You look at your phone. Will you press green to accept? Or red to decline?

Presumably, in this sort of situation you feel that at least two options are possible, or available for you to make. Even as you make your choice, you may feel that an alternative option is available. Additionally, you may feel that the choice is up to you – that you are, in some sense, the source of your choice. Such experiences of seemingly free agency, especially when they involve conscious deliberation, awareness of open possibilities, and a feeling that the choice is up to oneself, are often appealed to by philosophers and scientists discussing free will. These theorists draw on such experiences, and people's reports about them, in support of various positions about free will.

Philosophical debates about free will focus on a number of questions, including the question of whether free will is compatible with determinism (the *compatibility* question), as well as whether humans actually have free will (the *existence* question). Yet many of these questions depend crucially on a prior question about the concept of free will (the *conceptual* question). This question asks how free will is supposed to be defined or analyzed. Typically, the conceptual question is motivated and addressed from at least one of two angles: (a) our experiences of apparently free agency, and (b) the type of control required for morally responsible agency. Recently, philosophical debates about the compatibility and existence questions have

focused almost entirely on the latter angle, concerning the type of control needed for moral responsibility. Here, we will discuss instead the largely neglected angle of understanding the conceptual question in terms of our experiences of deliberation, choice, and action.

This angle has presumably been neglected in part because it is not always clear what the relevant experiences even are, or whether we can reliably introspect and describe them in such a way that they might provide information relevant to the conceptual question, or because it is unclear whether reliable information about experiences of free agency could justify certain answers to the conceptual question. Furthermore, even if consideration of experiences of agency helps to address some aspects of the conceptual question, it might not address aspects related to the type of control required for moral responsibility. And many philosophers are primarily motivated to understand the conditions of responsibility, i.e., for justified praise and blame, reward and punishment. Even so, experiences of free agency may be the more basic source of ordinary people's understanding of free will, and of their concerns about whether we have it. This issue has motivated scientists who are skeptical about whether we have free will (e.g., Wegner 2002; Harris 2012), since they take scientific discoveries to show that we lack the sort of free agency that we experience having (for responses to such claims, see, e.g., Mele 2009; Nahmias 2014).

In the next section, we lay some groundwork for thinking about experiences of agency, including free agency, and we outline some of the relevant literature on this topic. In subsequent sections, we focus on how claims about experiences of free agency relate to debates about free will. In these debates, incompatibilists argue that free will is inconsistent with determinism, whereas compatibilists argue that it is consistent. Libertarians are incompatibilists who think that we have free will, and so determinism is false. Typically, incompatibilists also think that our experience of free agency would be inaccurate if determinism is true. Call such experiences *libertarian*. By contrast, compatibilists think that experiences of free agency might be accurate, even assuming determinism. The accuracy of our experiences of free agency is related to the existence question, since it seems that such experiences are veridical only if we really are agents of the sort that we experience being.

The main question that we will address in this chapter is what to say about reportedly libertarian experiences of free agency – in other words, experiences of options as being open, and up to oneself to decide among, such that, if they are accurate or veridical, then (at a minimum) indeterminism must be true.² A great deal rides on this question. If normal experiences of free agency are libertarian, and if compatibilists cannot explain them away, then all of us may be under systematic illusion at almost every moment of our waking lives. That is, our experiences would be illusory if humans do not in fact satisfy libertarian conditions (e.g., if indeterminism does not occur in the right time or place during decision-making, or we lack agent-causal powers). In some sense, *more* rides on the question of whether we are agents of the sort that we experience ourselves as being than on the question of whether we have the type of control required for moral responsibility. After all, most philosophers agree that we can be morally responsible in a number of interesting senses even if determinism is true (see e.g., Shoemaker 2015 for an overview of these different senses of moral responsibility). Whatever disagreement exists is instead focused on a particular sense of moral responsibility – namely, a backward-looking, desert-based sense required specifically for retributive blame and punishment (see e.g., Pereboom 2014). At worst, we are not responsible in this sense if we lack what libertarians say we need. Yet even then, we would not be under systematic experiential illusion during most of our waking lives, as we are if our experience is libertarian, yet false.

2 Preliminaries

Part of the difficulty in thinking about free will is the tangle of issues that underlies any starting point. Typically, free will is taken to consist in an agent's possessing and exercising some cluster of executive or metaphysical control capacities. Sometimes, philosophers focus on these capacities in isolation – in other words, independently of how they might relate to an agent's moral responsibility for her behavior (e.g., Mele 1995, p. 3; Yao 2017). At other times, philosophers focus instead on whether such capacities are sufficient, in terms of control, for an agent to be morally responsible for what she does (e.g., McKenna 2012; Pereboom 2014). To complicate matters further, the relevant control capacities have often been understood to include the ability to do otherwise (Van Inwagen 1983; Vihvelin 2004; cf. Fara 2008), while at other times what is considered more important is whether an agent is the relevant source of her decision or action, even if she could *not* have done otherwise (e.g., Frankfurt 1969; Mele 1995; Fischer and Ravizza 1998; Pereboom 2014).

Although some work in metaphysics addresses questions about free will directly, rather than addressing them through an analysis of our concepts (e.g., Sartorio 2016), often what is of primary concern is how to analyze the concept of free will. Thus, the compatibility question, which asks whether free will is compatible with determinism, asks whether the concept of free will could refer successfully in a deterministic world. Similarly, the existence question, which asks whether we actually have free will, asks whether this concept refers in the actual world. Yet even here, various complex questions arise about the nature of this concept and how to think about its reference (e.g., Heller 1996; Murray and Nahmias 2012; Vargas 2013; Nichols 2013; Caruso 2015; Deery 2021a, 2021b). For instance, someone might think that the concept of free will refers only if our beliefs about free will are true, or not significantly in error. By contrast, others think that the concept might refer even if many of our beliefs about free will are false, or significantly erroneous.

Related to the conceptual question, at least on some versions of it, is the question of how we acquire the concept of free will (call this the *etiological* question). Focusing on this question, it seems plausible that we acquire the concept (a) as a result of tracking certain control capacities in other agents (often for the purpose of assigning responsibility), and (b) as a result of our own exercise of such capacities. Regarding the latter, one suggestion with a long pedigree is that we acquire the concept of free will partly due to how we *experience* our own seemingly free agency (see e.g., Strawson 1986). Yet even this apparently straightforward suggestion is, as it turns out, fraught with difficulty and raises a number of further questions.

For one thing, it is unclear what it even means to say that we *experience* our own agency, let alone that we experience *free* agency. According to some, what we call *experiences* of agency are really *beliefs* or *judgments* about the exercise of our own agency (e.g., Korsgaard 1996). On certain versions of this view, such “experiences” are really quasi-third-personal self-attributions of attitudes aimed at explaining, through rationalizing, our own behavior and sensory states, thereby fitting them into a “self-narrative” (Dennett 1992; Carruthers 2007). However, there are various reasons to doubt that such accounts can be the whole story (see Bayne and Pacherie 2007 for a detailed response to such views).

Moreover, even if we do experience (free) agency, a question arises about whether such experiences have proprietary phenomenology – i.e., experiential character that is different from any other type of experiential character – as visual or proprioceptive experiences do. Or

do they instead have non-proprietary phenomenology, perhaps like experiences of emotions? Relatedly, should we think that experiences of agency have *content*? Maybe agentic experiences are like experiences of tickles or itches, which some claim have no content – are not *about* anything – despite their having phenomenal character. And even if experiences of agency have content, what sort is it? Does it have to be explicitly representational, or might it be partly implicit, such that the overall content of the experience might outstrip what is explicitly represented in that experience (Horgan and Nichols 2016)?

We can also ask about the *structure* of agentic experiences. Do they have a world-to-mind direction of fit, or structure, like desires do (e.g., Searle 1983)? Or instead, do they have a mind-to-world direction of fit, like beliefs or perceptual experiences (see Bayne 2008 for discussion)? Assuming that the structure is like that of beliefs or perceptual experiences, so that experiences of agency have satisfaction conditions in the sense of having veridicality conditions (unlike desires, for instance, which have satisfaction conditions but *not* veridicality conditions), what might that content be? Here, a great deal depends on whether agentic experiences have liberal, or instead only sparse, content. In perceptual experience, liberal contents attribute not just low-level properties like redness or squareness, but also high-level properties like being an apple (e.g., Siegel 2009). Likewise, liberal contents for agentic experiences attribute high-level properties like acting freely, not just low-level properties like being an action (Bayne 2008 pp. 189–191).

In turn, the question of how liberal the content of agentic experiences is depends partly on what one takes to be the relation between content and phenomenology. According to *content-first* intentionalists, for instance, phenomenology is determined by intentional content (e.g., cf. Lycan 1996; Tye 2000; Carruthers 2000). As a result, it may seem reasonable to say that the content (if any) of experiences of agency is sparse (in fact, such theorists will most likely deny that there is, strictly speaking, any such phenomenology or content at all). By contrast, *phenomenology-first* intentionalists think that at least some intentional content is determined by phenomenology (cf. Siewert 1998; Horgan and Tienson 2002). As a result, it may seem natural to say that the content of our agentic experiences is liberal (e.g., cf. Siegel 2009; Horgan 2015; see Bayne 2008 for discussion). In particular, it is open to such theorists to talk in terms of *phenomenal content* (Kriegel 2002), which is content that is constitutively determined by phenomenology. If experiences of agency have phenomenal content, as some suggest, then this content might well be liberal enough to make the experiences count as experiences of *free* agency. Even so, the question remains whether such content is compatibilist or instead libertarian (i.e., veridical or not depending on whether determinism is true). Some maintain that our experience is too “anemic” (Nichols 2012, p. 293) to require the falsity of determinism for the experience to be accurate (cf. Bayne 2008, p. 196).

In what follows, we will assume that, normally, humans do experience their own agency, and these experiences can have liberal enough content to be experiences of *free* agency. We will, however, remain agnostic about whether experiences of agency have proprietary phenomenology, or whether instead they are experiences only in a non-proprietary sense – like emotions, for example. As a result, we will sometimes talk about experiences of free agency as having phenomenal content, in the way that they might if phenomenology determines content and such experiences have proprietary phenomenology, and at other times only of whether such experiences, more loosely construed, result in our making (for instance) certain judgments about free agency. We will also assume that experiences of free agency at least

partly explain how we acquire the concept of free will, although we will focus less on whether such experiences influence the application conditions of this concept – i.e., the conditions under which the concept correctly applies – or whether they support belief in the sort of free agency that we experience having.

With these assumptions in place, the main questions are, first, about the content of our experiences of free agency: are they libertarian or not? And second, if they *are* libertarian, do these experiences provide evidence for libertarian answers to the conceptual question or the existence question – i.e., do the experiences support a libertarian theory of free will? There is a long tradition of libertarians' claiming that our experience is full-blooded enough to have libertarian content, and that such experiences give us defeasible evidence for our having libertarian free will (e.g., Reid 1788, p. 36; O'Connor 1995, pp. 196–197; Swinburne 2012, p. 82; Guillon 2014). For instance, regarding the experience of being able to choose among open possibilities, John Searle asks us to:

[R]eflect very carefully on the character of the experiences you have as you engage in normal, everyday human actions. You will sense the possibility of alternative courses of action built into these experiences ... that we could be doing something else right here and now, that is, all other conditions remaining the same. This, I submit, is the source of our own unshakeable conviction of our own free will. (1984, p. 95)

Likewise, C. A. Campbell writes that

Everyone must make the introspective experiment for himself: but I may perhaps venture to report ... that I cannot help believing that it lies with me here and now, quite absolutely, which of two genuinely open possibilities I adopt. (1951, p. 463)

By contrast, some compatibilists insist that the relevant experiences have *obviously* compatibilist content, while others focus on the responsibility angle and remain silent on the issue of experience, perhaps because they assume that experiences of free agency are not robust enough to support a libertarian answer to the conceptual question.³

Yet, by rejecting the claim that experiences of free agency even *seem* to have libertarian content, these compatibilists risk ignoring one of the central motivations for incompatibilism, and they also risk offering analyses of free will that are inapt or incomplete. Ideally, of course, we could answer this question about the content of experiences of free agency by means of empirical evidence, rather than by relying on philosophers' competing introspective claims.

Unfortunately, however, the extremely limited empirical evidence that might bear on this question has not yet resolved the issue, since it appears to cut both ways. In some experiments, participants report libertarian experiences (Deery, Bedke, and Nichols 2013), while in others they report that their experience is compatibilist (Nahmias et al. 2004).⁴

In any case, to the extent that people report libertarian experiences for at least *some* of their choices, compatibilists incur an explanatory burden: they must explain libertarian reports about experiences of free agency. In other words, they must explain *away* the appearance of libertarian content in these experiences rather than simply deny that there even is such an appearance.

In the next three sections, we turn to recent compatibilist attempts to shoulder this burden. The first is developed by Oisín Deery (2015a, 2015b, 2021a). Deery assumes, at least for

argument's sake, that experiences of free agency have libertarian content. Still, such experiences also plausibly have a second sort of content that might be accurate, and thus compatibilist, assuming determinism. As a result, even experiences with libertarian content might be accurate if determinism is true, as long as the second sort of content is satisfied.

The second proposal is due to Terry Horgan (2007, 2011, 2012, 2015). Horgan grants that introspection *seems* to reveal that experiences of free agency are libertarian. Yet he insists that such introspection is not reliable. When people judge their experience of free agency as libertarian, they misinterpret it. Even when people think or say that their experience is libertarian, actually it is compatibilist.

Third, Deery (2015c, 2021a) develops an alternative view to Horgan's, which agrees with Horgan's assessment that experiences of free agency only seem to be libertarian, when in fact they are compatibilist. Yet Deery disagrees with Horgan regarding where the relevant mistake lies. For Deery, the mistake lies in how the experience itself is generated, rather than in how people interpret their experience.

3 Compatibilism about Libertarian Experience

According to Deery (2015a, 2015b, 2021a), even if we assume that experiences of free agency have libertarian phenomenal content, those very experiences might be veridical even if determinism is true. That is because these experiences might have more than one distinct type of phenomenal content, and the experiences might be veridical even if the satisfaction conditions of just one of these types of content are met.

Deery outlines how his suggestion works by analogy with a similar move that David Chalmers (2006) makes in connection with the phenomenal content of visual color experiences. Chalmers maintains that the view about phenomenal content that is most adequate to our normal phenomenology in color experiences is *primitivism*. According to this view, we experience colors visually as simple intrinsic properties of objects, spread out over their surfaces. As Chalmers explains it:

When I have a phenomenally red experience of an object, the object seems to be simply, primitively, *red*. The apparent redness does not seem to be a microphysical property, or a mental property, or a disposition, or an unspecified property that plays an appropriate causal role. Rather, it seems to be a simple qualitative property, with a distinctive sensuous nature. (2006, p. 66)

That is, experiences of color have contents that attribute primitive properties (cf. J. Campbell 1997). This is what Chalmers calls *perfect* content. He thinks it natural to judge such content as phenomenal content, given that the properties presented in the experience are constitutively determined by the phenomenology.

However, Chalmers points out that, "For all its virtues with respect to phenomenological adequacy, the ... primitivist view has a familiar problem. There is good reason to believe that the relevant primitive properties are not instantiated in our world" (2006, p. 66). If so, then none of our visual experiences of color is veridical. We are under an illusion at every moment of our waking lives (when we see colors).

In addition to perfect content, Chalmers defends another type of phenomenal content that makes color experiences veridical in the right kinds of cases. This is *imperfect* content, which

has its own associated veridicality condition: it is satisfied just in case the relevant object has whatever property (or set of properties) normally *causes* color experiences. The central idea is that experiences of color are (usually) veridical, despite the fact that primitive color properties are not actually instantiated.

Chalmers's view is complex. At bottom, however, imperfect phenomenal content is a *mode of presentation* of a property, rather than the property itself. When someone has an experience as of seeing a red apple, her experience attributes the property redness. For Chalmers, the second content is a mode of presentation of this property, which is "a condition that a property must satisfy in order to be the property attributed by the experience" (2006, p. 59).

Assuming that the property is a physical property, "one can naturally hold that the associated condition on this property is the following: it must be the property that normally causes phenomenally red experiences (in normal conditions for the perceiver)" (2006, p. 59). Such content is phenomenal since it is constitutively determined by the phenomenology.

For color experiences to be perfectly veridical, objects would have to instantiate primitive color properties. Yet even an experience that is not veridical in this way might be imperfectly veridical – i.e., veridical according to the standards by which we ordinarily differentiate veridical from non-veridical color experiences, as when we judge that we are seeing, rather than hallucinating, a red apple. For Chalmers, there is no conflict here, as long as we bear in mind that these two notions of veridicality are associated with distinct *conditions* of veridicality. A visual experience of color thus has more than one type of phenomenal content, depending on the associated notion of veridicality, and the experience counts as veridical as long as *one* of these conditions is satisfied.

Chalmers argues that the most fundamental type of content is perfect content. This is because what is presented in phenomenology determines the imperfect content via a "matching" relation, which works as follows. For a color experience to be perfectly veridical, we would have to live in "Eden," a world in which primitive color properties are instantiated (since that is what is presented in phenomenology). The best that we can do in our (presumably non-Edenic) world is to have certain properties "match" the primitive properties attributed by perfect content, by playing the role that those properties *would* play in Eden. Although no property can play this role perfectly, some property (or properties) may play it well enough, by being the normal cause of color experiences. In this way, imperfect phenomenal content is grounded in perfect content. In Chalmers' terms, perfect content serves as a "regulative ideal" in determining the imperfect content. Perfect content sets an ideal standard for the veridicality of phenomenal content, and the imperfect content is a condition that relates us to whatever properties come closest (in our world) to meeting that standard.

Deery (e.g., 2021a) argues that a similar story can be told for experiences of free agency. Even if we allow that what we are presented with in experience is libertarian, and thus is non-veridical if determinism is true, still there is a second type of phenomenal content that might be veridical, assuming determinism.

By analogy with primitivism about color experiences, it may seem reasonable to think that an experience of free agency with libertarian phenomenal content is veridical only if libertarianism is true. Deery calls a world in which libertarianism is true an "Agentive Eden." However, Deery maintains that experiences of free agency plausibly also have a second, imperfect, phenomenal content. Recall that this content is a condition that a property must satisfy in order to be the property that is attributed by the experience. Here, the attributed property is of various options being open for one to decide among, in a way that would require

the truth (minimally) of indeterminism. We feel that we could be “doing something else right here and now, ... all other conditions remaining the same” (Searle 1984, p. 95), or that that “it lies with me here and now, quite absolutely, which of two genuinely open possibilities I adopt (C. A. Campbell 1951, p. 463). What condition might work as the *imperfect* content for such an experience?

For color, the second content is whatever property (or set of properties) ordinarily causes phenomenal color experiences. In the agential case, the imperfect content might, Deery maintains, be the following condition: that there is instantiated whatever relevant property (or set of properties) is ordinarily instantiated when one experiences being free to decide among alternatives, or that causes such experiences. This content is veridical just in case this condition is satisfied, and there is no reason to think that it could not be satisfied if determinism is true. Moreover, the content is phenomenal since it is constitutively determined by the phenomenology.

Deery argues that among the two types of phenomenal content, the most fundamental is perfect content. That is because we are assuming that perfect content most accurately reflects what is presented in phenomenology, which is that we are free in the way described by libertarians. Analogously with Chalmers’s view, Deery maintains that this libertarian content determines the second, imperfect, content, via a matching relation. For an experience of free agency to be perfectly veridical – i.e., veridical according to the standards associated with its perfect content – we would have to live in an Agentive Eden. The best that we can do if determinism is true (for example) is to have certain properties match the libertarian properties that are attributed by the perfect content, by playing the role that these properties *would* play in an Agentive Eden. No property can play this role perfectly. Yet some property (or set of properties) may be able to play it well enough, by being the property (or set of properties) that ordinarily *causes* experiences of free agency – e.g., when choosing among various options. This content is compatibilist: it might be veridical even if determinism is true, since it depends only on what *actually* underlies or causes such experiences.⁵ Nevertheless, this compatibilist imperfect content is grounded, Deery maintains, in libertarian perfect content, since (to use Chalmers’s phrase) the perfect content acts as a “regulative ideal” in determining the imperfect content. In other words, perfect content sets an ideal standard for veridicality, and the imperfect content is a condition that relates us to whatever properties come closest (assuming determinism) to meeting that ideal standard. Once the second content is satisfied, the experience is imperfectly veridical, even if determinism is true – and this despite granting that the relevant phenomenology is libertarian.

Thus, an experience of free agency can be veridical, it can have libertarian phenomenal content, yet libertarianism can be false (i.e., we might not be libertarian agents). The mistake, according to Deery, is to presuppose that there is a *unique* phenomenal content to experiences of free agency – namely, libertarian perfect content. Plausibly, as with our color experiences, such experiences also have an imperfect phenomenal content that is compatibilist. If determinism is true then, even granting (for the sake of argument) that experiences of free agency have libertarian perfect content, there is still a sense in which *these very experiences* might be veridical – despite their veridicality not being dependent on their libertarian content. On this view, libertarian experiences of free will might provide reasons to *believe* free will requires libertarian conditions, yet compatibilism need not be an *error theory*, and it might be true, including for the type of free will that grounds moral responsibility.

4 Horgan's View: Error Theory #1

Let us dispense now with the assumption that experiences of free agency have libertarian content, and consider two error theories for introspective reports that suggest that they *do* have such content.

First, Terry Horgan (2007, 2011, 2012, 2015) grants that introspection may *seem* to reveal libertarian content. Nevertheless, Horgan insists that introspection is not reliable in this domain. Horgan agrees that people often *judge* their experience of free agency to be libertarian. However, he argues that when people do so, they misinterpret their experiences, which do not actually have content that would be illusory if determinism were true. By spelling out how this happens, Horgan gives an error theory for libertarian judgments about experience.

Horgan begins by conceding that people often judge their experience to be libertarian. For instance, he allows that

... when one attends introspectively to one's free-agency phenomenology, with its presentational aspect of ... freedom ... and when one simultaneously asks reflectively whether the veridicality of this phenomenology requires ... libertarianism, one feels some tendency to judge that the answer to this question is Yes. (2011, p. 94; cf., 2007, p. 23)

However, Horgan thinks that while introspection is reliable in some domains, introspective judgments about whether one's experience of free agency is libertarian are unreliable. He begins by distinguishing between two sorts of introspection: (a) attentive introspection, which involves "paying attention to certain aspects of one's current experience" (Horgan 2011, p. 84), and (b) judgmental introspection, "the process of forming a judgment about the nature of one's current experience" (2011, p. 84). The content that we attentively introspect is "presentational content," which is "the kind that accrues to phenomenology directly – apart from whether or not one has the capacity to articulate this content linguistically and understand what one is thus articulating" (2011, p. 91).

In judgmental introspection, by contrast, we attend to certain aspects of our experience, but we also form judgments about those aspects. Thus, "Judgmental introspection ... deploys attentive introspection, while also generating a judgment about what is being attended to" (2011, p. 84). For Horgan, there cannot be an appearance/reality gap when we attentively introspect. Yet in judgmentally introspecting, we can go wrong: we might be subject to what Horgan calls a "labeling fallacy" (2012, p. 408–409). For instance, we might make a performance error in applying the concept 'red' to our experience of redness: we might mistakenly apply the concept 'green.' In Horgan's parlance, we might "mislabel" the phenomenology. Presumably, this sort of mistake hardly (if) ever happens with simple sensory experiences. As a result, while attentive introspection is infallible, judgmental introspection is not quite infallible, although it typically is, especially with simple experiences.

Even so, Horgan claims, our judgments about whether our experiences of free agency are libertarian (or compatibilist, for that matter) are highly fallible. For a start, answering this question goes beyond what attentive introspection is capable of: we cannot read off the answer from phenomenology. The question can only be answered by judgmental introspection. Yet Horgan thinks that when we try to answer the question of whether our experience is libertarian (or instead compatibilist) by judgmentally introspecting, we find that we cannot arrive at a reliable

answer, even though the question is about the character of our own introspectively available experiences. It is not just that we are subject to the occasional labeling fallacy.

Horgan thinks that there is a good explanation for this inability, as we will outline in a moment. Yet he admits that many philosophers tend to judge their experience of free agency as libertarian, and says, “I confess to experiencing some temptation to think so myself ... a temptation that needs explaining” (2012, p. 416). To this end, Horgan offers a two-part debunking explanation for such judgments. First, he suggests that if we think we can tell by introspection that our experience is libertarian, this may reflect a form of “introspective con-fabulation.” It is one thing to know (a) by introspection:

- a) My experience does not present my behavior as determined by my prior states.

Yet it is another thing to know (b) by introspecting on phenomenology:

- b) My experience presents my behavior as not determined by my prior states.
(Cf. Horgan 2015, p. 54)

Horgan admits that we can ascertain whether (a) is true by introspecting. However, (b) is distinct from (a), and we cannot ascertain whether (b) is true by introspection. Even if (b) were true, we could not know this by judgmentally introspecting. When we judge our experience to be libertarian, and thereby assert (b), either we are mistakenly inferring (b) from (a), or simply conflating (a) and (b).

Additionally, Horgan thinks that the concept of free will has compatibilist application conditions, and the veridicality conditions of experiences of free agency “coincide” with these application conditions. This claim is important since it bears on the second part of Horgan’s two-part debunking explanation of libertarian judgments about experiences of free agency, as we will now explain.

Part of why Horgan thinks that the concept of free will is compatibilist is that people appear to be competent in applying this concept in ordinary contexts – for instance, when they distinguish free from unfree choices or actions (e.g., where agents are coerced at gunpoint, or are subject to irresistible addictions, and so on). Compatibilism accommodates these judgments easily, by enabling them to come out true even under the assumption of determinism. By contrast, libertarians require that a more stringent condition be met, namely, that indeterminism (at a minimum) be true. Horgan thinks that we should prefer compatibilism to libertarianism (and to incompatibilism more generally) since, all else being equal, one hypothesis – compatibilism – is better than another – libertarianism – if it accommodates how competent users of the relevant concept ordinarily apply it.

This issue is important for the second part of Horgan’s two-part explanation of libertarian judgments about experience. Here, Horgan tells a contextualist story about the application conditions of the concept of free will, which also applies to judgments about *experiences* of free agency. Horgan maintains that “the very posing of the question whether human freedom is compatible with ... determinism tends to alter the contextually operative settings on certain implicit semantic parameters that govern the concept *freedom* – and tends to drive those parameter settings so high that, in the newly created context, no item of behavior that is ... determined counts as *free*” (Horgan 2007, p. 22).

Horgan grants that contextual parameters of this sort do not plausibly apply (directly, at least) to experiences of free agency. After all, he thinks that many non-human animals share

with us “a fair amount of agential phenomenology” (2007, p. 10), despite the fact that their mental content is not governed by contextual semantic parameters. Even so, when we introspect on our experiences of free agency while also asking ourselves whether they are veridical if determinism is true, Horgan thinks that our introspective *judgment* about the experience gets “infected” by the same confusion that occurs whenever we ask the compatibility question about determinism and the *concept* of free will. On Horgan’s view, it is understandable why our experiences might lead people to believe in libertarian free will, but these beliefs can be explained away so that they do not provide evidence against compatibilism.

5 Prospection and Causal Modeling: Error Theory #2

Deery (2015c, 2021a) agrees with Horgan that to judge an experience of free agency as libertarian is to make a mistake. Yet Deery disagrees with Horgan about where the mistake lies. For Deery, the mistake lies in how the experience itself is generated, rather than in how people interpret the experience.

According to Deery, people judge their experiences of free agency as libertarian because the relevant experiences are generated by prospection, which is the mental simulation of future possibilities for the purpose of guiding action. When experiences of prospection are understood in terms of causal modeling, the result is a mechanism by which the experience itself seems libertarian, even though it is not.

Prospection, as Martin Seligman, Peter Railton, Roy Baumeister, and Chandra Sripada (2013) outline it, “is guidance... by present, evaluative representations of possible future states. These representations can be understood minimally as ‘If X, then Y’ conditionals, and the process of prospection can be understood as the generation and evaluation of these conditionals” (2013, p. 119). For Seligman and colleagues, agents – in order to regulate their interactions with the environment – construct representational mental models of that environment. The most efficient models will be of the form, “if in circumstance C and state S, then behavior B has outcome O with probability p” (2013, p. 124). Such “feedforward/feedback” models will typically have the following type of structure:

expectation → observation → discrepancy detection → discrepancy-reducing change in expectation → expectation → ...

On this picture, agents generate and use simulations of future possibilities, often by drawing on and learning from past experience, and the function of these simulations is to enable the agents to navigate effectively into the future by selecting suitable actions. Prospection thus nicely captures the forward-looking character of typical experiences of free agency, since agents experience their options as a “branching array of evaluative prospects that fan out before them” (Seligman et al. 2013, p. 119).

Most prospection occurs outside of conscious awareness and is unavailable to introspection, since it would be inefficient for agents to consciously keep track of all the simulations that they generate. Yet prospection is sometimes consciously experienced. According to Seligman and colleagues, *affect* plays a central role in its becoming conscious. When prospection encounters “incommensurable dimensions and conflicting values and perspectives”

(2013, p. 131), explicit comparison of these factors is facilitated by the brain's "common metric" of affect, such that "conscious subjective affect attached to prospecting ... enable[s] them to compete effectively with ongoing experience" (131). Thus, when agents have conflicting thoughts about what to do, their simulated options feed into "an experientially rich and detailed workspace," with the result that agents can "use their intelligence and imagination to best effect" (2013, p. 131). In such cases, "it can be best to act in awareness of ... conflicting thoughts" (131). As a result, prospection is consciously experienced (cf. Sripada 2016; Nahmias 2016).

Deery (e.g., 2021a) maintains that prospection explains how agents can experience mere future possibilities for choice. Yet it fails to explain why this sort of experience seems indeterministic, and thus libertarian. Deery maintains that causal modeling provides the missing, seemingly indeterministic element.

According to Deery, it is natural to interpret the hypotheticals generated in prospection as carrying causal information about what *would* happen under variations in the values of variables – alternative choices the agent might make – in a causal model. A causal model is a representation that encodes hypothetical relationships between variables, which represent causal relata (i.e., events). Evidence suggests that ordinary causal cognition is indeed underpinned by such modeling (Sloman 2005; Lagnado, Gerstenberg, and Zultan 2013).

In causal modeling, to establish whether one event, X, causally influences another event, Y, we consider what would happen to Y by altering X's value (for full details, see Woodward 2003). If a change reliably occurs in the value of Y, then X is judged to be a cause of Y. In intervening on X in this way, causal modeling requires that we ignore the prior causal variables that normally result in X's value, and instead we allow X to vary freely across a range of alternative values, which it could not otherwise take. In other words, we treat it as an exogenous variable, i.e., as a variable whose values are determined by factors outside the model, rather than as an endogenous variable whose values are determined by the values of other variables within the model.

Philosophers have recently begun to use this sort of modeling to illuminate various questions about free will (e.g., Roskies 2012; Deery and Nahmias 2017). For instance, Jenann Ismael (2013) maintains that agents mentally construct models of this sort when deliberating about what to do. In choosing among options for action, an agent carves off the event of her choice from its actual causal antecedents, and treats it as an exogenous variable in a causal model. By doing so, she is enabled to assess the "downstream" effects of this variable's varying across a range of values, which yields causal information relevant to action-planning.

These are the very hypotheticals that prospection generates in regulating the agent's interaction with the environment.

Deery maintains that prospection – together with a causal modeling account of how the hypotheticals generated by prospection should be modeled – explains why people end up having experiences of free agency that *seem* libertarian. First, when agents generate simulated possibilities for action while deliberating about what to do, the variable representing their choice is treated by prospection as a free variable, meaning that it is permitted to vary over a range of values. Yet, were the deliberating agent to consider the same choice while assuming determinism, she would instead treat it as having antecedent sufficient causes, and therefore as a variable the values of which are constrained by the wider model of the deterministic system.

In that case, the variable is permitted to take just a single value. This creates an apparent psychological conflict between treating one and the same variable as both free and

constrained. When an agent tries to hold in mind both models of her decision at the same time – for instance, in a forced choice experiment in which she is asked whether her experience of free agency is consistent with determinism (e.g., as in Deery, Bedke, and Nichols 2013) – each model might be experienced as inconsistent with the other. As a result, the experience of one's choice as possibly taking any of several values might seem inconsistent with determinism, which would permit the choice to take just a single value.

However, while these two models may be experienced as inconsistent, they are not. As Ismael puts it, “there is no more conflict between these models than there is between the view of a building from close-up and the view from a very great distance” (2013, p. 230). In prospection, we see our choice “from close up,” by modeling it as an exogenous variable. Yet, when we are asked to model that choice within a wider deterministic system, we see it “from a very great distance,” since we treat it as an endogenous variable within that system. Thus, even if an experience of free agency *seems* inconsistent with determinism – and therefore libertarian – due to the felt inconsistency of the two models, it does not follow that the experience is libertarian, since these models are just different ways of modeling the choice, each of which may be useful for different purposes.

Deery also outlines a second reason why experiences of free agency might seem libertarian. When prospection models a choice, that choice appears more open – perhaps even indeterministically open – than when it is modeled as an endogenous variable, for instance as part of a deterministic system. Deery explains this increased sense of openness in terms of *rich* epistemic possibility.

To maintain that it is epistemically open whether you will choose either of two options is simply to maintain that your choosing either option is consistent with what you know (cf. Kapitan 1986; Pereboom 2008, p. 292–296). Famously, J.J.C. Smart (1961) maintained that this is how we often interpret counterfactuals outside the sphere of action. For instance, when we say, “the plate fell, and it could have broken,” we are not making a claim about determinism. All we are saying is that before the plate completed its fall, for all we knew it would break (1961, p. 298). Similarly, if we say that Lee Harvey Oswald could have done otherwise than shoot President Kennedy, we are saying that before Oswald pulled the trigger, for all we knew he would not. In making claims about epistemic possibility, there is clearly no conflict with determinism. As a result, epistemic possibility seems a poor candidate for explaining the sense of openness that Deery seeks to explain.

However, Deery thinks that the rich epistemic possibility involved in deliberation and choice is liable to be interpreted as inconsistent with determinism – and hence as libertarian – since an agent's prospection ignores the choice's prior causes, by treating it as an *exogenous* variable in a causal model. Prospection thus ignores a large part of what the agent actually knows, or might reasonably be expected to bring to mind in other contexts – e.g., that her choice *has* prior (perhaps even sufficient) causes. After all, when we deliberate, we typically think about the *effects* of the various choices that we are considering, and we do not think about all of the causes that might lead us to choose one among various options (much less about the causes of the considerations that come to mind during deliberation). As a result, the epistemic possibilities that are available to the agent in prospection have a restriction on the “for all I know” that yields ordinary epistemic possibility, making the possibilities feel “richer.” This leads the agent to experience her available possibilities for choice as more robustly open than they would otherwise appear to be – even to the point of their seeming (at least implicitly) to be indeterministically open.

On Deery's view, therefore, the suggestion of indeterministic or libertarian openness is built right into ordinary experiences of free agency, as a result of the rich epistemic possibility at work in the causal modeling that *generates* the experiences in prospect. Nevertheless, this suggestiveness does not amount to anything like libertarian content, since the possibilities remain epistemic, and thus entirely consistent with determinism. Furthermore, while the agent is – in a sense – misinterpreting her experience when she judges it as indeterministically open or libertarian, matters are not as straightforward as Horgan makes them out to be when he claims that there is nothing at all in the experiences (as attentively introspected) that is suggestive of such indeterministic openness. Instead, this error theory explains why theorists believe our experiences are libertarian and provide (defeasible) evidence for libertarian free will, but also why they are mistaken since the experiences are consistent with compatibilist theories of free will.

6 Conclusion

Some philosophers and scientists appeal to experiences of free agency in support of various positions about free will. Skeptics about free will argue that these experiences suggest conditions that we humans fail to meet, and as a result we live under systematic illusion. Some libertarians argue that such experiences provide evidence that we actually possess libertarian free will. These moves require that (a) we have experiences of free agency and (b) these experiences have content that we tend to introspect as having libertarian veridicality conditions (e.g., indeterminism). Compatibilists *can* argue that their accounts of free will do not conflict with our experiences of free agency by rejecting either one of these two requirements. Here, we have considered what compatibilists might say if they do *not* reject them, but instead concede that experiences of free agency appear to have libertarian content. Even while making this concession, compatibilists can argue that libertarian content has compatibilist veridicality conditions, or they can provide an error theory to explain why our experience of free will might *seem* to have libertarian features, even though, in fact, they do not.

Notes

- 1 Deery is primary author. For helpful comments we would like to thank Joe Campbell and Terry Horgan.
- 2 In what follows we will often use 'determinism' to stand in for a more general claim about whatever theses conflict with the conditions required for libertarianism, which at a minimum include indeterminism, but typically indeterministically caused events at specific times and places during decision-making (e.g., Kane 1996) and often more metaphysically robust conditions such as agent-causal powers (e.g., O'Connor 1995). Hard incompatibilists or skeptics about free will agree that these libertarian conditions are required for free will, but believe that humans do not satisfy these conditions. Some skeptics motivate the need for libertarian conditions with claims about the experience of free agency or will (e.g., Harris 2012), while others focus on the requirements for desert (e.g., Pereboom 2014). Very few argue that determinism is true; rather, they argue that agent-causation is implausible, that a more general thesis such as physicalism is plausible and rules out libertarianism, or that libertarian conditions are metaphysically impossible or incoherent (e.g., Strawson 1986).

- 3 For representative compatibilist views of this sort, see e.g., Grünbaum (1952), who writes:

Let us carefully examine the content of the feeling that on a certain occasion we could have acted other than the way we did... Does the feeling we have inform us that we could have acted otherwise under exactly the same external and internal motivational conditions? No, ... this feeling simply discloses that we were able to act in accord with our strongest desire at that time, and that we could indeed have acted otherwise if a different motive had prevailed at the time. (1952, p. 672)

More recently, compatibilists have offered analyses of the ability to do or choose otherwise that are compatible with determinism, but without focusing on whether these analyses accurately capture our experiences of choice (e.g., Vihvelin 2004; Fara 2008).

- 4 There is also some debate about what sort of experiences should be picked out as paradigmatic experiences of free agency. Libertarians sometimes focus on “close call” or “torn” choices, for which the reasons are closely balanced even at the moment of choice (Campbell 1951; Strawson 1986; Balaguer 2010), perhaps because such choices seem to allow a role for indeterministic events (van Inwagen 1989; Kane 1996). Some compatibilists, however, suggest that such indecisive choices are experienced as relatively unfree. Instead, paradigm experiences of free agency occur when the agent deliberates to reach a *confident* (rather than a torn) decision about what to do (Nahmias 2006; Lau, Hiemisch, and Baumeister 2015).
- 5 Deery (2015a, 2021a) argues that there is good reason to think that this condition constitutes a genuine compatibilist content for experiences of free agency. He does so by analogy with the view that *FREE CHOICE* is a natural-kind concept that refers to whatever relevant capacities agents actually exercise when (under normal conditions) they make paradigmatically free choices (cf. Heller 1996). On this view, it is irrelevant to free agency whether determinism is true. We choose freely, unless the relevant capacities fail to constitute a relevant kind. Consequently, we might be free even if determinism is true. Such views are widely held about concepts like *WATER* (e.g., Putnam 1975).

Bibliography

- Balaguer, M. (2010). *Free Will as an Open Scientific Problem*. Cambridge, MA: MIT Press.
- Bayne, T. (2008). The phenomenology of agency. *Philosophy Compass* 3 (1): 182–202.
- Bayne, T. and Pacherie, E. (2007). Narrators and comparators: The architecture of agential self-awareness. *Synthese* 159: 475–491.
- Campbell, C.A. (1951). Is “Freewill” a pseudo-problem? *Mind* 60 (240): 441–465.
- Campbell, J. (1997). The simple view of colour. In: *Readings on Color* (ed. A. Byrne and D. Hilbert), 257–268. Cambridge, MA: The MIT Press.
- Caruso, G. (2015). Free will eliminativism: Reference, error, and phenomenology. *Philosophical Studies* 172 (110): 2823–2033.
- Carruthers, P. (2000). *Phenomenal Consciousness: A Naturalistic Theory*. Cambridge University Press.
- Carruthers, P. (2007). The illusion of conscious will. *Synthese* 159: 197–213.
- Chalmers, D. (2006). Perception and the fall from Eden. In: *Perceptual Experience* (ed. T. Gendler and J. Hawthorne), 49–125. Oxford: Oxford University Press.
- Deery, O. (2015a). The fall from Eden: Why libertarianism isn’t justified by experience. *Australasian Journal of Philosophy* 93 (2): 319–334.
- Deery, O. (2015b). Is agential experience compatible with determinism? *Philosophical Explorations* 18 (1): 2–19.
- Deery, O. (2015c). Why people believe in indeterminist free will. *Philosophical Studies* 172 (8): 2033–2054.
- Deery, O. (2021a). *Naturally Free Action*. New York: Oxford University Press.

- Deery, O. (2021b). Free actions as a natural kind. *Synthese* 198: 823–843.
- Deery, O., Bedke, M., and Nichols, S. (2013). Phenomenal abilities: Incompatibilism and the experience of agency. In: *Oxford Studies in Agency and Responsibility, Volume 1* (ed. D. Shoemaker), 126–150. Oxford: Oxford University Press.
- Deery, O. and Nahmias, E. (2017). Defeating manipulation arguments: Interventionist causation and compatibilist sourcehood. *Philosophical Studies*. 174 (5):1255–1276.
- Dennett, D. (1992). The self as a center of narrative gravity. In: *Self and Consciousness: Multiple Perspectives* (ed. F. Kessel, P. Cole and D. Johnson). Hillsdale, NJ: Erlbaum.
- Fara, M. (2008). Masked abilities and compatibilism. *Mind* 117 (468): 843–865.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
- Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66 (23): 829–839.
- Guillon, J.B. (2014). Van inwagen on introspected freedom. *Philosophical Studies* 168: 645–63.
- Grünbaum, A. (1952). Causality and the science of human behavior. *American Scientist* 40 (4): 665–676.
- Harris, S. (2012). *Free Will*. New York: Free Press.
- Heller, M. (1996). The mad scientist meets the robot cats: Compatibilism, kinds, and counterexamples. *Philosophy and Phenomenological Research* 56 (2): 333–337.
- Horgan, T. (2007). Agentive phenomenal intentionality and the limits of introspection. *Psyche* 13: 1–29.
- Horgan, T. (2011). The phenomenology of agency and freedom: Lessons from introspection and lessons from its limits. *Humana Mente* 15: 77–97.
- Horgan, T. (2012). Introspection about phenomenal consciousness: Running the gamut from infallibility to impotence. In: *Introspection and Consciousness* (ed. D. Smithies and D. Stoljar), 403–422. Oxford: Oxford University Press.
- Horgan, T. (2015). Injecting the phenomenology of free will into the free will debate. In: *Oxford Studies in Agency and Responsibility, Volume 3* (ed. D. Shoemaker), 34–61. Oxford: Oxford University Press.
- Horgan, T. and Nichols, S. (2016). The Zero Point and I. In: *Pre-Reflective Consciousness: Sartre and Contemporary Philosophy of Mind* (Eds. S. Miguens, G. Preyer and C. Bravo Morando), 143–175. Taylor and Francis.
- Horgan, T. and Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In: *Philosophy of Mind: Classical and Contemporary Readings* (ed. D. Chalmers), 520–532. Oxford: Oxford University Press.
- Ismael, J. (2013). Causation, free will, and naturalism. In: *Scientific Metaphysics* (ed. H. Kincaid, J. Ladyman and D. Ross), 208–235. New York: Oxford University Press.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.
- Kapitan, T. (1986). Deliberation and the presumption of open alternatives. *The Philosophical Quarterly* 36 (143): 230–251.
- Korsgaard, C. (1996). *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.
- Kriegel, U. (2002). Phenomenal Content. *Erkenntnis* 57: 175–198.
- Lagnado, D., Gerstenberg, T., and Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science* 37: 1036–1073.
- Lau, S., Hiemisch, A., and Baumeister, R. (2015). The experience of freedom in decisions – questioning philosophical beliefs in favor of psychological determinants. *Consciousness and Cognition* 33: 30–46.
- Lycan, W. (1996). *Consciousness and Experience*. Cambridge: MIT Press.
- Mele, A. (1995). *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press.
- Mele, A. (2009). *Effective Intentions: The Power of Conscious Will*. New York: Oxford University Press.
- McKenna, M. (2012). *Conversation and Responsibility*. New York: Oxford University Press.

- Murray, D. and Nahmias, E. (2012). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, 88 (2): 434–467.
- Nahmias, E. (2016). Free will as a psychological accomplishment. In: *The Oxford Handbook of Freedom* (ed. D. Schmitz and C. Pavel), 492–507. New York: Oxford University Press.
- Nahmias, E. (2006). Close calls and the confident agent: free will, deliberation, and alternative possibilities. *Philosophical Studies* 131 (3): 627–667.
- Nahmias, E. (2014). Is free will an illusion? Confronting challenges from the modern mind sciences. In: *Moral Psychology, Vol. 4: Freedom and Responsibility* (ed. W. Sinnott-Armstrong), 1–25. Cambridge, MA: MIT Press.
- Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2004). The phenomenology of free will. *Journal of Consciousness Studies* 11 (7–8): 162–179.
- Nichols, S. (2012) The indeterminist intuition: Source and status. *The Monist* 95 (2): 290–307.
- Nichols, S. (2013). Free will and error. In: *Exploring the Illusion of Free Will and Moral Responsibility* (ed. G. Caruso), 203–218. Lanham, MA: Lexington.
- O'Connor, T. (1995). Agent causation. In: *Agents, Causes, and Events: Essays on Indeterminism and Free Will* (ed. T. O'Connor), 173–200. New York: Oxford University Press.
- Pereboom, D. (2008). A compatibilist account of the epistemic conditions on rational deliberation. *The Journal of Ethics* 12 (3): 287–306.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
- Putnam, H. (1975). The meaning of “Meaning”. *Minnesota Studies in the Philosophy of Science* 7: 131–193.
- Reid, T. (1788). *Essays on the Active Powers of Man*. Edinburgh: Bell & Robinson.
- Roskies, A. (2012). Don’t panic: Self-authorship without obscure metaphysics. *Philosophical Perspectives* 26 (1): 323–342.
- Sartorio, C. (2016). *Causation and Free Will*. Oxford: Oxford University Press.
- Searle, J. (1983). *Intentionality*. Cambridge, MA: Cambridge University Press.
- Searle, J. (1984). *Minds, Brains, and Science*. Cambridge, MA: Harvard University Press.
- Seligman, M., Railton, P., Baumeister, R., and Sripada, C. (2013). Navigating into the future or driven by the past: Prospection as an organizing principle of mind. *Perspectives on Psychological Science* 8 (2): 119–141.
- Shoemaker, D. (2015). *Responsibility from the Margins*. Oxford: Oxford University Press.
- Siegel, S. (2009). The visual experience of causation. *The Philosophical Quarterly* 59 (236): 519–540.
- Siewert, C. (1998). *The Significance of Consciousness*. Princeton: Princeton University Press.
- Sloman, S. (2005). *Causal Models: How People Think about the World and Its Alternatives*. New York: Oxford University Press.
- Smart, J.J.C. (1961). Free will, praise and blame. *Mind* 70: 291–306.
- Sripada, C. (2016). Free will and the construction of options. *Philosophical Studies*. 10.1007/s11098-016-0643-1.
- Strawson, G. (1986). *Freedom and Belief*. Oxford: Clarendon Press.
- Swinburne, R. (2012). *Free Will and Science*. London: British Academy.
- Tye, M. (2000). *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.
- Van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Oxford University Press.
- Van Inwagen, P. (1989). When is the will free? *Philosophical Perspectives* 3: 399–422.
- Vargas, M. (2013). If free will doesn’t exist, neither does water. In: *Exploring the Illusion of Free Will and Moral Responsibility* (ed. G. Caruso), 177–202. Lanham, MA: Lexington.
- Vihvelin, K. (2004). Free will demystified: a dispositional account. *Philosophical Topics* 32 (1–2): 427–450.
- Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.
- Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.
- Yao, V. (2017). Strong-Willed Akrasia. In: *Oxford Studies in Agency and Responsibility, Volume 4* (ed. D. Shoemaker), 6–17. Oxford: Oxford University Press.

The Future of the Causal Quest

HANNAH TIERNEY

1 Introduction

In one sense, questions about the nature of free will can be understood as ethical or normative questions. We arguably care about being free to the extent that we care about being morally responsible. In this way, the nature of free will is inextricably tied to the nature of moral responsibility. But questions about the nature of free will can also be understood as metaphysical questions. Though many take free will to be the kind of freedom or control required to be morally responsible, it's understood as a necessary metaphysical condition for moral responsibility.¹ Given that free will is a metaphysical concept, it stands to reason that work on other metaphysical concepts can shed light on the nature of free will. And work on the metaphysics of laws, dispositions, and abilities has certainly impacted and informed our understanding of free will in a meaningful way. However, there has been a relative reluctance to turn to the literature on the metaphysics of causation. This reluctance is perhaps best captured by Peter van Inwagen, who famously wrote: "Causation is a morass in which I for one refuse to set foot. Or not unless I am pushed" (van Inwagen 1983, p. 65).² But this appears to be changing, for several philosophers have recently begun to wade into the morass.³ And as we'll see, though it's becoming increasingly agreed upon that causation and free will are related, it's much less clear how we should go about incorporating the metaphysics of causation into our accounts of free will. In this chapter, I will look at three recent attempts to draw lessons about free will from the causation literature: Oisín Deery and Eddy Nahmias's (2017) account of interventionist causation and manipulation arguments, Carolina Sartorio's (2016) actual causal sequence account of free will, and Sara Bernstein's (2017) analysis of the relationship between causal proportionality and moral responsibility. What follows is far from an exhaustive analysis of the current work on causation and free will and in focusing on these particular views I've ignored many others. What I find compelling about these particular views is that though they represent three very different ways of incorporating work on causation into discussions of free will, they all face real challenges about how best to conceive of the relationship between the metaphysical and ethical questions regarding the nature of free will. And by reflecting on

A Companion to Free Will, First Edition. Edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White.
© 2023 John Wiley & Sons, Inc. Published 2023 by John Wiley & Sons, Inc.

the different ways those working on free will can utilize the research on causation, and on the questions about the interplay between metaphysics and ethics that these approaches raise, we can reveal new and interesting avenues for future research not only on the relationship between causation and free will but on the metaphysics of free will more generally.

2 Deery and Nahmias's Interventionist Response to Manipulation Arguments

In their recent paper, "Defeating Manipulation Arguments: Interventionist Causation and Compatibilist Sourcehood," Oisín Deery and Eddy Nahmias (2017) use an interventionist account of causation to defend compatibilists against one of the most serious objections on the market: manipulation arguments.⁴ There are many different kinds of manipulation arguments, but the most familiar instances share a common structure, which Michael McKenna has helpfully laid out:

- (1) If S is manipulated in manner X to A, then S does not A of her own free will and is therefore not morally responsible for A'ing.
- (2) An agent manipulated in manner X to A is no different in any relevant respect from any normally functioning agent determined to do A from (CAS) [the compatibilist-friendly agential structure].
- (3) Therefore, if S is a normally functioning agent determined to A from CAS, she does not A on her own free will and therefore is not morally responsible for A'ing (McKenna 2008, p. 143).

Deery and Nahmias use a case drawn from Alfred Mele's manipulation argument (2013) to develop their response.

First, imagine Danny. One evening in 1986, Danny's parents made love, hoping to conceive a child. They got lucky. A zygote was formed (at time t_1), and nine months later Danny was born. Thirty years later, Danny is walking down a deserted street and he finds a wallet with the owner's ID in it and \$500. Danny takes himself to have good reasons for keeping the money, but also for returning the wallet. He deliberates for a while, and in the end he decides to keep the money, and he does so (at time t_{30}). Assume that this occurs in a *deterministic* universe – that is, a universe in which, for each event E, the laws of nature and some set of events that occurred prior to E are such that these events cause E to occur with probability 1. If determinism is true, then some set of events prior to Danny's act of stealing the wallet at t_{30} are (together with the laws) such that they cause his deliberating and acting in that way, at that time, with probability 1. (Deery and Nahmias 2017, p. 1257)

Compare this to a different case:

[A] powerful Goddess, Diana, has the power to know what will happen in the future and to act in ways that ensure that specific events occur in the distant future. Diana has these abilities in part because she exists in a deterministic universe and is able to get enough information about events occurring in it (e.g., at t_1) to deduce exactly what she needs to do at that time to ensure that a particular event occurs thirty years later. In this case, Diana assembles atoms in a specific way at t_1 so as to create a zygote that develops into a child, grows up, finds a wallet thirty years later, and at t_{30} decides to keep the money it contains. For some reason, Diana wants to ensure that this event occurs at t_{30} , and she possesses the power to alter events at t_1 precisely so that she ensures that it does occur.

As it turns out, the life of this intentionally created person (whom we will call Manny since he is *Manipulated*) follows the exact same course as the life of deterministic Danny (as described above). Manny is no different from Danny when it comes to his abilities to consider options, weigh reasons, and make decisions about whether to steal the money. (Deery and Nahmias 2017, p. 1257)

The manipulation argument can proceed as follows. Intuitively, Manny does not have free will and is not morally responsible for stealing the money, and more generally, any agent who is manipulated in the way that Manny is does not have free will. Deery and Nahmias call this the NoFW Premise. But there is no principled difference between Manny and Danny – they meet all the same compatibilist sufficient conditions for free will – and there is no difference between the kind of manipulation featured in the case of Manny and the truth of determinism when it comes to free will and moral responsibility.⁵ Deery and Nahmias call this the NoDif premise. So, Danny doesn't have free will and isn't morally responsible for stealing the money, just as no agent has free will and is morally responsible if determinism is true.

There are many different avenues a compatibilist can take in responding to manipulation arguments. McKenna calls the strategy of denying the NoFW Premise taking the “hard-line” while those who deny the NoDif premise take the “soft-line” (McKenna 2008). McKenna favors taking the hard-line, since it appears that soft-line responses are mere stop-gap solutions that only keep the incompatibilist at bay temporarily. A compatibilist can point to a condition for free will and moral responsibility that the manipulated agent does not meet that the determined agent does, but it seems in principle possible for the defender of the manipulation argument to augment the cases such that the manipulated agent meets the condition in question as well. But while taking the soft-line can (arguably) only succeed temporarily, taking the hard-line is no easy path; it requires the compatibilist to accept the (arguably) counterintuitive claim that Manny and those manipulated like him do in fact have free will and are morally responsible for the actions for which they were manipulated to perform.

Deery and Nahmias argue that it's possible for the soft-line to be more than a mere stop-gap defense in the face of manipulation arguments. In their recent essay, they reject the NoDif Premise and argue that there is an in principle difference between Manny (and those who are manipulated like him) and Danny (and those who exist in a world where determinism is true). They also contend that there is no way (or easy way) for the defender of manipulation arguments to alter their case to sure up the NoDif Premise in response to their soft-line defense of compatibilism. Deery and Nahmias rely on interventionist approaches to causation to reject the NoDif Premise.⁶ Interventionism has a growing influence in the causation literature, and has figured in several recent discussions of agency and free will.⁷ Deery and Nahmias argue that it can also be used to develop an account of the minimally sufficient conditions for free will that are immune to manipulation arguments.

Interventionism about causation involves constructing models that represent counterfactual relationships among events. We first assign variables to the event-types in question, and we can set these variables to different values in order to represent particular event-tokens (Woodward 2003). If we want to know whether the output of Danny (or Manny's) CAS caused him to steal the money, we first assign variables to the output of his CAS (X) and the stealing of the money (Y). Next, we can perform an intervention to determine whether X caused Y , which involves setting X to a different value and seeing whether such an intervention reliably changes the value of Y . For example, we can imagine that the output of Danny's CAS was the decision not to steal the money. In this case, it seems likely that this intervention would reliably result in a change in the value of Y , i.e. the stealing of the money wouldn't occur. We can also

make finer-grained interventions, where we imagine that the output of Danny's CAS is the decision to steal the money but takes on a slightly weaker motivational force. Given this intervention, the value of Y will likely be different as well – perhaps Danny will delay in stealing the money so the event of stealing the money takes place at a slightly different time.

On an interventionist approach to causation, the output of Danny (and Manny's) CAS counts as a direct cause of Danny (and Manny) stealing the money. This is because an intervention on X would reliably change the value of Y . However, being a direct cause of an event is distinct from being the *causal source* of the event, and according to many, causal sourcehood is what we're after when we seek to determine moral responsibility, praise, and blame (Cartwright 1979). According to Deery and Nahmias, for a cause to be the causal source of a given event, that cause must bear the strongest causal invariance relation to it among all the events that are causally relevant (Deery and Nahmias 2017). Deery and Nahmias characterize causal invariance as follows:

A causal invariance relation, R_1 , that obtains between two causal variables, X and Y , is stronger than another such relation, R_2 , obtaining between Y and another of its prior causal variables – for instance, W – if:

- (1) holding fixed the relevant background conditions, C , R_1 predicts the value of Y under a wider range of interventions on X than R_2 does under interventions on W ; and
- (2) R_1 predicts the value of Y across a wider range of relevant changes to the values of C than R_2 does. (Deery and Nahmias 2017, pp. 1262–1263)

Deery and Nahmias argue that in order for an agent to be free and thus morally responsible for a given action, the output of an agent's CAS must be the causal source of that action, in that it must bear the strongest causal invariance relation to that action. In the case of Danny, his CAS plausibly is the causal source of his decision to steal the money. We've already seen how changes in the value of the variable representing the output of Danny's CAS result in a change to the value of the variable representing Danny's stealing the money. And because Danny is an intentional agent, his decision to steal the wallet would lead to him to do so across many changes in background conditions. Danny would likely steal the money if it was in a slightly different location, if it were raining instead of snowing, if the wallet was black instead of brown, etc. Of course, there will be some changes in the background conditions that would result in Danny not stealing the money, i.e. if a police officer was standing nearby.

This is all true of Manny as well. In the case of both Danny and Manny, the relationship between the variables representing the outputs of their CAS and the variables representing their stealing the money is equally strong. But there is a stronger causal invariance relation between Diana's meddling and Manny's stealing, according to Deery and Nahmias.

The relation between Diana's decision and Manny's action is such that, across a maximally wide range of changes to the background conditions, the variable representing Manny's stealing does not change in value without a change in the value of the variable representing Diana's decision, and changes in the value of the variable representing Diana's decision correspondingly change the value of the variable representing Manny's decision to steal. (Deery and Nahmias 2017, p. 1264)

Importantly, this is because Diana intends for Manny to steal the money at t_{30} and she is able to ensure that he does so. If Diana is able to ensure that Manny steals the money at t_{30} , then there's

no change in the background circumstances that could alter the value of the variable representing Manny stealing the money. When evaluating the relationship between the output of Manny's CAS and his stealing the money, we must ignore Diana and all other causally relevant events – we are only concerned with the output of Manny's CAS and his stealing the money. And just as there are changes to the background circumstances that would stop Danny from stealing the money, i.e. if a police officer was standing nearby, these changes would stop Manny as well. Thus, Diana's decision is the source of Manny stealing the wallet, not the output of his CAS.

From here, Deery and Nahmias are able to develop a soft-line response to the manipulation argument. They have isolated a relevant difference between Danny and Manny – while the output of Danny's CAS is the causal source of his stealing the money, the output of Manny's CAS is not. Thus, they can reject the NoDif premise both as it pertains to Danny and Manny and the more general claim that there is no difference between the kind of manipulation featured in Manny's case and the truth of determinism when it comes to free will. In this way, Deery and Nahmias have put forth a soft-line strategy that is meant to be much more than a stop-gap response to manipulation arguments, for they have isolated an in principle difference between manipulation and determinism,⁸ while the former renders agents unable to be the causal sources of their actions, the latter does no such thing.

There is much to say about Deery and Nahmias's innovative approach to manipulation arguments. First, their use of an interventionist approach to causation is emblematic of a current trend in the work on free will. John Campbell (2010) uses interventionism to explain a notion of mental causation, Adina Roskies (2012) uses it to construct an account of self-authorship, and Jenann Ismael (2013, 2016) relies on it to defuse the threat of determinism to free will. As work on interventionist approaches to causation continues to advance, the extent to which these views can be utilized to illuminate the nature free will will surely be a topic of future exploration.

On the one hand, interventionism seems to be a particularly good approach for those who work on free will. Recall that the question of whether we have free will can be understood as both a metaphysical and ethical question, where the free will or control condition is typically understood as the metaphysical condition necessary for agents to be the appropriate targets of our judgments of moral responsibility, praise, and blame. In other words, free will is a metaphysical relation beholden to our normative practices. Interventionism approaches causation in a similar way. For instance, interventionist approaches require us to choose which variables are endogenous and exogenous in modeling causal relationships (Ismael 2016). If we think that an agent's intentions or the output of their CAS is causally relevant to a given outcome, we can assign independent variables to these features (as opposed to others) in constructing our model. In this way, our interests and values are essential features of interventionist causal models, for they are key to determining how systems are to be modeled. But interventionist approaches to causation are neither purely normative nor subjective. As Ismael explains:

Which networks we are interested in, and which variables we treat as endogenous and exogenous, are determined by context and purpose ... [C]ausal relations ... are inductive generalizations from testable regularities and are grounded in the fact that the world is built from a collection of relatively autonomous rule-governed components. (Ismael 2016, p. 136)

Because of the success interventionism has found in other areas of both philosophy and science, and because the interventionist approach to causation is able to bridge the formidable gap between our practices and metaphysics, many working on free will find this approach to be a fruitful and illuminating source of inspiration.

On the other hand, some philosophers are skeptical that interventionism can be used to inform accounts of free will. Sara Bernstein (2017), for example, argues that because interventionism relies on normative practices and human judgment, it provides an account of *causal explanation* but not *causation*, which is a mind-independent feature of the world (Bernstein 2017, p. 178). This can be understood as both an objection to interventionist accounts of causation in general and an objection to their use in developing accounts of free will in particular. The general worry is that interventionist accounts are only able to give us an account of our causal judgments or explanations, but this is importantly different from giving us an account of causation. But there's also a particular worry about the use of interventionist approaches to causation in accounts of free will: if we want the answer to the question "what is a cause?" to inform the answer to the question "what is free will?" we don't want our intuitions about the latter question to affect the answer to the former. This would be, if not entirely circular, quite uninformative.

Deery and Nahmias are aware of this objection, at least in its general form. They argue:

[O]bjections claiming that interventionism gives us, at best, an account of our casual practices or judgments, rather than providing (as it should) a theory of what metaphysically causes what, come close to begging an important methodological question ... Surely it is an open methodological question...whether the interventionist approach to causation – as well as to other topics in metaphysics – is the correct one to adopt ...We do not assume that it is, yet we insist that it is a strong contender, which, if correct, supports a soft-line response to the Manipulation Argument. (Deery and Nahmias 2017, p. 1270)

Notice that Deery and Nahmias can be right about the methodological point as to whether interventionism counts as an account of causation, but this doesn't address the more particular objection to using interventionist accounts of causation to inform our accounts of free will. While there might be nothing dialectically improper about relying on our intuitions about free will to generate an account of causation, there does seem to be something illicit about relying on our intuitions about free will to generate an account of causation that will then go on to inform an account of free will.

An objection to Deery and Nahmias's soft-line objection may help to illustrate this criticism.⁹ Deery and Nahmias consider several objections to their view, including the objection that their causal sourcehood requirement is too demanding. In many cases, there could be causal variables outside of an agent's CAS that could bear a more invariant relation to an agent's action than the output of an agent's CAS. In such cases, the output of an agent's CAS would not be the causal source of the agent's action (or at least not the only causal source). Deery and Nahmias take this to be a positive feature of their view, for it allows them to make sense of free will that comes in degrees and our scalar practices of holding agents morally responsible.¹⁰ In such cases, the agent might still be free and morally responsible for the action in question, though not maximally responsible.

I think that Deery and Nahmias's response to this objection moves too quickly. Diana is the causal source of Manny's stealing the wallet because she *acts through* his compatibilist agential structure, ensuring that he steals the wallet. But our compatibilist agential structures are often acted through, without eliminating or mitigating the degree to which we are free and responsible. We often perform blameworthy actions because others ask us to, or put the idea in our head, or give us an incentive to do so. These kinds of causal variables act through our compatibilist agential structures, and while they don't ensure that we'll perform these blameworthy actions, they do make it more likely. In these kinds of cases, Deery and Nahmias's analysis seems to commit them to the claim that these external causal variables are

the causal sources of our blameworthy behavior, thus counterintuitively eliminating (or at least reducing) the degree to which the agents who performed the blameworthy actions are free and responsible.

Imagine the following case: A child, who loves ponies very much, asks his mother to get him a pony. The mother, like most parents, doesn't have the financial resources to buy a pony but because she loves her son very much and her son wants a pony very badly, she decides to steal a pony for him. Now imagine that the mother acts on her decision and steals a pony. In this case, it strikes me that the mother acted freely in stealing the pony and is morally responsible for doing so. But is she the causal source of the action?

Using Deery and Nahmias's analysis, we can model the relationship between the son's request for a pony (variable R , value r), the mother's decision to steal a pony (variable D , value d), and the mother stealing the pony (variable S , value s).

$$\begin{aligned} R=r &\rightarrow D=d \rightarrow S=s \\ D=d &\rightarrow S=s \end{aligned}$$

If we hold the relevant background conditions, C , fixed, interventions on both R and D will produce similar changes to the value of S . If the son never asked for a pony, the thought of stealing a pony would never have occurred to the mother, and she wouldn't have stolen the pony. And if the mother had never decided to steal the pony, then she wouldn't have committed the crime. But if we consider a range of changes in C , R bears a stronger invariance relation to S than D bears to S . Because the son asks his mother to give him a pony, we can predict that she would go through with stealing a pony across more changes to the circumstances than if she had simply decided to steal a pony to give to her son. Perhaps the mother would be willing to climb a higher fence, or undergo the crime even in bad weather if her child asked her to get him a pony, but she wouldn't do so if she simply decided to steal a pony for her son without being asked. Recall that when evaluating the relationship between D and S , we must ignore the contribution of R . So, for all the mother knows, her son might not even want a pony! Thus, the son's request for a pony (R) bears a stronger invariance relation to the mother's action of stealing the pony (S) than the mother's decision alone (D) bears to the event. And according to Deery and Nahmias's account, this means that the son's request for a pony is the causal source of the mother's action of stealing the pony, not her decision to do so. But surely the mother is morally responsible for stealing the pony, even if the causal source of the action was her child's request. And while Deery and Nahmias leave open the possibility that this can simply mitigate, as opposed to eliminate, the degree to which the mother is free and morally responsible, it's odd that the mother's degree of moral responsibility is *at all* mitigated simply because she was responding to a child's request. Surely stealing a pony because your child asks you for a pony is just as blameworthy as stealing a pony because you know your child likes ponies.

One natural response to this objection is to argue that in the case described above, the mother's compatibilist agential structure encompasses the fact that her child asked her for a pony. After all, many accounts of free will rely on notions of reasons-responsiveness (Fischer and Ravizza 1998) and reasons sensitivity (Sartorio 2016), so it would make sense to include the reason to engage in the blameworthy behavior as a feature of the agent's compatibilist agential structure. Other accounts rely on first- and second-order desires (Frankfurt 1971), and perhaps the child's request for a pony gives rise to a first-order desire to steal a pony that is in-line with a second-order desire to desire things that will make one's child happy. And if this is the case, then even when we're doing an intervention on D and

ignoring the role of *R*, we still must take into account that the child asked the mother for a pony (either as a reason for acting or as a desire to be fulfilled), in which case *D* would bear at least as invariant a relation to *S* as *R* would. In this case, *D* really would be the causal source of *S*, and Deery and Nahmias could argue that the mother is fully responsible for stealing the pony.

But notice that this response relies on using other views of free will to determine which variables are endogenous to the model. But the very reason we constructed the model was to answer the question: “Was the mother free to, and morally responsible for, stealing the pony?” While there’s nothing illegitimate about using views of free will to assign variables in an interventionist model when the question we’re trying to answer is about causation, there is something dangerously circular and uninformative about using a conception of free will to determine which variables to represent within an interventionist model when the question we’re trying to answer is about free will.

Of course, the above is a very preliminary worry and there’s much to say in Deery and Nahmias’s defense. Here I only want to argue that as the influence of interventionist models of causation grows, a fruitful topic of research will be how to incorporate interventionism into accounts of free will in an informative and illuminating way.

3 Sartorio’s Actual Causal Sequence View

In her recent book *Causation & Free Will* (2016), Carolina Sartorio develops the Actual Causal Sequence view of free will, or ACS. Like Deery and Nahmias, Sartorio takes causation to play a central role in free will. However, unlike Deery and Nahmias, Sartorio doesn’t commit to a single account of causation.¹¹ Rather, Sartorio identifies key features of causation (or a relevantly similar metaphysical relation) and uses them to develop ACS. There isn’t enough room to do justice to Sartorio’s rich and innovative account of free will. In this chapter, I’d like to focus on one feature in particular – the interplay between the inspiration for the view – Frankfurt cases – and the commitments that come with the features of causation Sartorio uses to develop her view.

Like others who develop actual-sequence views,¹² Sartorio takes the inspiration of ACS to be Frankfurt cases. Frankfurt cases, originally developed by Harry Frankfurt (1969), are designed to illustrate the irrelevance of the ability to do otherwise and the importance of an agent’s actual sequence in accounting for free will and moral responsibility. Sartorio develops her own Frankfurt case as follows:

Frankfurt Case: A neuroscientist has been secretly monitoring the brain processes of an agent, call him Frank, who is deliberating about whether to make a certain choice, *C*. The neuroscientist can reliably predict the choices that Frank is about to make by looking at the activity in his brain, and can also manipulate Frank’s brain in a way that guarantees that Frank will make choice *C*. He plans to intervene if he predicts that Frank will not make choice *C* on his own. As it happens, Frank makes choice *C* on his own, motivated by his own reasons, and without the intervention of the neuroscientist (who correctly predicts that Frank would make that choice on his own). (Sartorio 2016, p. 13)

Sartorio isolates two intuitions that are generated by the above Frankfurt case:

Intuition 1: Frank (our agent in a Frankfurt case) is in control of his act despite his lack of robust alternatives.

Intuition 2: What determines whether Frank is in control of his act is how he actually came to perform the act. (Sartorio 2016, p. 17)

Intuition 1 isolates what *isn't* relevant to free will and control while Intuition 2 isolates what *is*. The fact that, intuitively, Frank is free and morally responsible for making choice C though he couldn't have done other than make choice C counts against views that defend the principle of alternative possibilities. The intuition that the actual sequence of events that lead Frank to choose choice C is what makes him free and morally responsible lends support to actual-sequence approaches to free will. Of course, an incredible amount has been written on whether these intuitions can either successfully undermine (in the case of Intuition 1) or support (in the case of Intuition 2) the relevant philosophical theses. Sartorio doesn't engage in this debate, but her development of the ACS view is inspired by these intuitions (Sartorio 2016, p. 16).

Sartorio develops two grounding claims that actual-sequence theorists are committed to in virtue of taking Frankfurt cases seriously. She then provides causal interpretations of both:

Positive grounding claim: freedom is grounded in facts about actual causal histories. (Sartorio 2016, p. 21)

Negative grounding claim: freedom isn't grounded in anything other than actual causal histories. (Sartorio 2016, pp. 28–29)

Sartorio then argues that these two claims together support a further claim about supervenience: “An agent's freedom with respect to X supervenes on those elements of the causal sequence issuing in X that ground that agent's freedom” (Sartorio 2016, p. 29). According to Sartorio, this claim captures the key insight from Frankfurt cases and actual-sequence theorists of free will should remain loyal to it in developing their views.

Sartorio then goes on to isolate several key features of causation (or a relevantly similar metaphysical relation) that can bolster the grounding and supervenience claims of ACS. In this chapter, I'd like to focus on two of those features:

OMISSIONS: omissions and other kinds of absences can enter into causal relationships. (Sartorio 2016, p. 46)

EXTRINSICNESS: a causal relation between C and E may obtain, in part, owing to factors that are extrinsic to the causal process linking C and E. (Sartorio 2016, p. 71)

On Sartorio's account, we can be responsible for our omissions, just as we can be responsible for our actions,¹³ and whether we cause these omissions and actions can depend on extrinsic factors. EXTRINSICNESS and OMISSIONS interact with each other in interesting ways and help defend ACS against counterexamples to the supervenience claim. Take this pair of cases originally discussed by van Inwagen (1983):

Phones: I witness a man being robbed and beaten. I consider calling the police. I could easily pick up the phone and call them. But I decide against it, out of a combination of fear and laziness.

No Phones: Everything is the same as in Phones except that, unbeknownst to me, I couldn't have called the police (the phone lines were down at the time). (Sartorio 2016, p. 56)

Intuitively, we are morally responsible in Phones but not in No Phones. Some, like van Inwagen (1983), argue that this is because we have the ability to do otherwise in Phones but not in No Phones. If van Inwagen is right, this would threaten the supervenience claim Sartorio,

and other actual-sequence theorists, defend. But Sartorio is able to explain the asymmetry in terms of the extrinsicness of causation.

In *Phones*, it's clear that the combination of fear and laziness caused the failure to call the police (because we're assuming that omissions can be caused), according to Sartorio (2016, p. 69). In *No Phones*, the combination of fear and laziness, though it caused the failure to *try* to call the police, it didn't cause the failure to call the police. Rather, an extrinsic factor – the fact that the phone lines were down – makes it impossible that the combination of fear and laziness could have caused the failure to call the police (Sartorio 2016, p. 90). An extrinsic factor is also involved in the causal story to be told in *Phones*. The combination of fear and laziness caused the failure to call the police in part because the phone lines *were* working (Sartorio 2016, p. 87). Notice that extrinsic features determine the causal relationships that obtain in both *Phones* and *No Phones*, and thus they determine the extent to which the agents in these cases are free (and responsible), but both of these features are out of the agents' control – they are a matter of luck. Sartorio calls this kind of luck that arises from the extrinsicness of causation “Type-2 luck.”

Type-2 Luck: The agent is not in control of some facts external to the causal history that help determine its composition. (Sartorio 2016, p. 89)

Interestingly, omissions and other kinds of absences are more susceptible to type-2 luck than actions and other positive events; Sartorio calls this Type-2 Luck Asymmetry (Sartorio 2016, p. 91). This is because extrinsic features are more likely to affect the causal history of omissions than they do actions. To illustrate this asymmetry, Sartorio compares *No Phones* to a Frankfurt-style version of *Phones*:

Frankfurt-style Phones [Action]: Again, I witness the man being robbed and beaten. This time the phones are working. With a lot of effort I manage to overcome my fear and laziness, pick up the phone and call the police. Unbeknownst to me, a neuroscientist has been monitoring my brain. Had I wavered in my decision, he would have manipulated my brain in such a way that I would still have made the same choice. (Sartorio 2016, p. 90)

Both *Phones* and Frankfurt-style *Phones* [Action] feature extrinsic factors, but these features don't affect the causal history in these cases symmetrically. The fact that the phone lines are down in *Phones* is an extrinsic factor that makes a causal difference – because the phone lines are down, the combination of fear and laziness doesn't cause the failure to call the police. In Frankfurt-style *Phones* [Action], the neuroscientist is an extrinsic factor that doesn't make a causal difference. Overcoming the fear and laziness is a cause of the agent calling the police, regardless of whether a neuroscientist stands by to intervene.

This asymmetry raises many questions, namely: how does ACS handle Frankfurt-style cases that involve omissions? Indeed, Sartorio considers such a case:

Imagine that I decide on my own not to call the police in circumstances where the phones were working; however, had I hesitated in making that choice, a neuroscientist who had been monitoring my thoughts would have intervened by manipulating my brain and, as a result, I would have made the same choice. Am I free and responsible for not calling the police in this case? (Sartorio 2016, p. 91)

It strikes me that the agent in the above case was free to, and responsible for, not calling the police, just as Frank is morally responsible in Sartorio's original Frankfurt case. But if one is committed to Type-2 Luck Asymmetry, this is far from clear. Rather than discuss how ACS would address Frankfurt-style cases featuring omissions, Sartorio moves past the issue, citing the vast literature on this very contentious topic,¹⁴ and arguing that it's unclear how generalizable the asymmetry between omissions and actions is.¹⁵

Sartorio's discussion of Type-2 Luck Asymmetry brings out an interesting tension between the motivation for ACS and the metaphysical assumptions on which it relies. The main, if not sole, motivation for ACS (and presumably other actual-sequence views) is the set of intuitions generated by Frankfurt cases. As Sartorio argues, these intuitions can be captured by two grounding claims, one positive and one negative, which in turn support a claim about supervenience: "No difference in freedom without a difference in the relevant elements of the causal sequence" (Sartorio 2016, p. 32). Sartorio then defends the supervenience claim from purported counterexamples by relying on the extrinsicness of causation. However, it is this very metaphysical assumption that makes it difficult for ACS to capture our intuitions when it comes to Frankfurt cases that feature omissions. But why should our judgments about a Frankfurt case featuring an omission be beholden to our metaphysical commitments? Why can't they be treated as the motivation for an actual sequence view, like the judgments about Frankfurt cases featuring actions?

We can alter Sartorio's original Frankfurt case to feature an omission:

Frankfurt Case-Omission: A neuroscientist has been secretly monitoring the brain processes of an agent, call him Frank, who is deliberating about whether to make a certain choice, C. The neuroscientist can reliably predict the choices that Frank is about to make by looking at the activity in his brain, and can also manipulate Frank's brain in a way that guarantees that Frank will **not** make choice C. He plans to intervene if he predicts that Frank will make choice C on his own. As it happens, Frank does **not** make choice C on his own, motivated by his own reasons, and without the intervention of the neuroscientist (who correctly predicts that Frank would **not** make that choice on his own).

I take it that this case can generate the intuition that Frank is morally responsible for not choosing choice C just as clearly and forcefully as Sartorio's original case can generate the intuition that Frank is morally responsible for choosing choice C. Notice that our intuitions in response to this pair of cases can still be captured by both Sartorio's positive and negative grounding claims, which can still support the supervenience claim – it can still be true that there's no difference in freedom without a difference in the causal sequences. Of course, we would need to rely on very different features of causation (or a metaphysically similar relation) to defend the supervenience claim in the face of purported counterexamples. Though we would most assuredly hold onto OMISSIONS, we might not be able to rely on EXTRINSICNESS, for example.

Whether we should conceive of Frankfurt cases featuring actions and Frankfurt cases featuring omissions as playing the same dialectical role is an interesting question. While some Frankfurt cases featuring omissions generate intuitions similar to the intuitions generated by Frankfurt cases featuring actions (like Frankfurt Case-Omission), others arguably do not. Indeed, some rely on this intuitive asymmetry to object to actual sequence views of free will (Swenson 2015, 2016). What generates these varying intuitions when it comes to

omissions? Is it possible for an actual sequence theorist to come up with an in principle difference between cases like Frankfurt Case-Omission and No Phones? And, more generally, when should we let our intuitions about freedom guide our metaphysical commitments and when should we let metaphysical commitments guide our judgments about freedom? These are difficult questions, and while many philosophers are currently grappling with them now, they are likely to become even more important as the focus on the metaphysics of causation in the free will literature grows more prominent.

4 Bernstein's Principle of Proportionality

So far I've reviewed two different approaches to the relationship between free will and the metaphysics of causation. Deery and Nahmias utilized a specific view of causation, the interventionist approach, to develop an account of free will that can withstand manipulation objections. Sartorio, rather than utilize a specific view of causation, relied on several common properties of causation to support ACS. Though both approaches are innovative and serve to move the research on free will forward, they each raise interesting and perplexing questions about the relationship between our normative and metaphysical commitments surrounding causation and free will. The use of interventionist causation when attempting to discover the causal source of agents' actions can either produce counterintuitive results or be uninformative, depending on the way in which we conceive of agents' compatibilist agential structures and represent them in our causal models. And while EXTRINSICNESS might be a plausible feature of causation, it leads to counterintuitive results when it comes to Frankfurt cases featuring omissions. While the above authors clearly agree that free will and causation are intimately connected, it's less clear how we can use theories of causation to illuminate the nature of free will. Sara Bernstein drives this worry home in her essay "Causal Proportions and Moral Responsibility" (2017).

Bernstein begins with the intuition that we can only be morally responsible for what we cause. Given this intuition, if one thinks that moral responsibility (or free will) can come in degrees, then it seems the following principle is true:

Proportionality: An agent's moral responsibility for an outcome is proportionate to her actual causal contribution to the outcome. (Bernstein 2017, p. 167)

On the face of it, this claim is much weaker than the views the authors above defend and can be accepted even by those who defend a variety of views about the relationship between free will and causation, but it too proves difficult for views of free will to fully accommodate. While this principle tells us that agents can be more or less responsible for outcomes, it alone cannot tell us when an agent is more or less responsible than another agent. To answer this question, we would need to know when one agent causally contributes to an outcome to a greater degree than another. But this cannot easily be accounted for by current theories of causation. First, Bernstein argues that depending on whether you favor a productive or dependent theory of causation (Hall 2004), an agent could causally contribute to an outcome to a greater or lesser degree and thus be either more or less responsible than another agent (Bernstein 2017). According to Bernstein, there is a "semantic indeterminacy" at play in Proportionality and it's not clear which theory of causation we should use to resolve it.¹⁶ Next, Bernstein argues that to truly understand Proportionality, we ought to have a precise metaphysical account of what it means for an agent to be more or less of a cause. But again it's

not clear that any theory of causation can give us such an account. Bernstein argues: “[T]he relationship between causation and moral responsibility so often used in moral assessment is much trickier than previously imagined. It is also particularly methodologically fraught if current causal theories are to be the guides” (Bernstein 2017, p. 181).

I won’t delve into the intricacies of Bernstein’s argument here. I would like to highlight how difficult the task of incorporating the metaphysics of causation into accounts of free will threatens to be given Bernstein’s analysis. In the sections above, I pointed to potential worries involved in relying on interventionist approaches to causation and the extrinsicness of causation. But Bernstein argues that there is *no* (single) theory of causation (or set of metaphysical assumptions about causation) that is able to successfully illuminate the nature of free will.¹⁷

Bernstein concludes her essay by arguing: “There is much work to be done before we can trust that linking moral responsibility to metaphysical theories of causation clarifies our theories rather than obfuscates our thinking on these matters” (Bernstein 2017, p. 181). But what kind of work is required and who should be given this task, those who work on causation or those who work on free will? Should those developing accounts of free will wait to incorporate theories of causation until there is consensus on what causation amounts to? This doesn’t seem particularly promising. Nor does it seem advisable to entirely ignore the role of causation in developing accounts of free will (if you take causation to be relevant to free will). But what work can be done in the absence of a single, determinate theory of causation that can perfectly account for the entirety of our intuitions regarding causation and moral responsibility?

One tentative answer is that those who work on free will can get clear on exactly the kind of work they want a theory of causation to accomplish. Take the two cases Bernstein originally compares in her essay:

Victim: Two independently employed assassins, unaware of each other, are dispatched to eliminate Victim. Being struck by one bullet is sufficient to kill Victim. Each assassin shoots, and Victim dies.

Hardy Victim: Two independently employed assassins, each unaware of the other, are dispatched to eliminate Victim. Unbeknownst to both assassins, Victim is particularly hardy, and requires two bullets for his demise. Each assassin shoots, and Victim dies. (Bernstein 2017, p. 165)

The first is a case of overdetermination, and the second is a case of joint causation. One can grant that different theories of causation will provide different assessments of causal contribution in these cases, and that there are no rules for which theory of causation we should use to determine who is more causally responsible, as Bernstein (2017) argues. But whether this is a problem depends entirely on what we want from a theory of causation.

I don’t have a clear intuition as to who is more morally responsible for killing Victim, either the assassins in Victim or Hardy Victim. Or rather, I have conflicting intuitions – in one sense the assassins in Victim seem more morally responsible and in another the assassins in Hardy Victim seem more responsible. And the two theories of causation that Bernstein discusses are able to accommodate these conflicting intuitions: on a counterfactual approach, the assassins in Hardy Victim are more causally responsible (and, given Proportionality, morally responsible) and on a productive approach to causation, the assassins in Victim are more causally responsible (and thus morally responsible). If what we want from a theory of causation is to be able to accommodate, explain, or perhaps even ground our judgments of moral responsibility, then we have a success.

But if what we want is for a single theory of causation to be able to resolve all outlier and problem cases of free will and moral responsibility, then we have a failure. But why would we

ever expect a single theory of causation to be able to do that? There are outlier and problem cases in the causation literature as well, and if no theory can render coherent all of our causal intuitions, then it's unreasonable to expect such a theory to render coherent all of our intuitions about moral responsibility. Of course, Bernstein's conclusion stands: there's much work to do. But those who work on free will can get their hands dirty right alongside the metaphysicians.

First, Bernstein argues that there are "no clear, principled rules for which type of causal relation should be used" to evaluate cases like Victim and Hardy Victim (2017, p. 172). Perhaps the literature on free will and moral responsibility could be helpful in developing such rules (if we want such rules in the first place). While there might be nothing in the causation literature that would lead one to favor one approach to causation over another, there might be normative concerns that could favor one approach. Second, there might be room for both (or more) notions of causation in our accounts of free will and moral responsibility.¹⁸ Perhaps we can be pluralists about causation – different notions of causation can ground different notions of free will and/or responsibility, for example.¹⁹ Of course, there are many ways of being a pluralist about causation, free will, and moral responsibility, and it's far beyond the scope of the chapter to discuss the details of such views. I only want to suggest here that the development of such views might prove to be an important focus of future research.

At the very least, we should adopt a methodological pluralism about causation. After all, it turns out that those working on free will want very different things from an account of causation. Deery and Nahmias want an account of causation that can make sense of the intuition that manipulated agents aren't free and responsible while determined agents are. Sartorio wants an account of causation upon which free will can supervene. And Bernstein seeks an account of causation that could come in degrees and explain our scalar judgments of moral responsibility. It's unlikely that a single conception of causation can accomplish all of these tasks, and perhaps we shouldn't expect it to. Rather, we can make real headway into the nature of free will by adopting many different approaches to causation, at least in the absence of a consensus about the true, single nature of causation.

5 Conclusion

In the concluding section of this chapter, I'd first like to summarize some of the questions that the current work on causation and free will raise:

- When utilizing an interventionist approach to causation in an account of free will, how can we assign variables to our causal models in a way that does not produce counterintuitive results or run the risk of providing circular/uninformative accounts of free will?
- Should those who defend actual sequence accounts of free will be committed to symmetrical judgments in the face of Frankfurt cases featuring omissions and Frankfurt cases featuring actions?
- What work can be done when it comes to free will in the absence of a single, determinate theory of causation that can perfectly account for the entirety of our intuitions regarding causation and moral responsibility?

And here are a few questions that can guide future research:

- In developing accounts of free will that involve causation, must philosophers defend a particular view of causation (like Deery and Nahmias), or simply rely on a set of plausible metaphysical assumptions about causation (like Sartorio)?

- When should we let our intuitions about freedom guide our commitments regarding causation, and when should we let these commitments guide our judgments about freedom?
- Can those who work on free will embrace a pluralism about causation in a fruitful way?

Finally, though I focused solely on the relationship between causation and free will in this chapter, notice that these guiding questions for future research apply to all areas in which metaphysics intersects with free will.

- In developing accounts of free will that involve metaphysical concepts, must philosophers defend a particular view of that concept, or simply rely on a set of plausible metaphysical assumptions about it?
- When should we let our intuitions about freedom guide our metaphysical commitments and when should we let these commitments guide our judgments about freedom?
- Can those who work on free will embrace a pluralism about any/all relevant metaphysical concepts in a fruitful way?

Notes

- 1 Often contrasted with an epistemic condition, which refers to the kind of knowledge or awareness required to be morally responsible.
- 2 Those who defend agent-causal libertarian views (e.g. O'Connor 1995) and event-causal libertarian views (e.g. Kane 1996) of free will are notable exceptions.
- 3 For example, Dana Nelkin (2011) has developed a compatibilist account of agent causation and Kadri Vihvelin (2013) relies on the metaphysics of causation in defending her dispositionalist account of free will.
- 4 For another interventionist response to the manipulation argument, see Usher (2020).
- 5 See Frankfurt (1971), Wolf (1990), and Fischer and Ravizza (1998) for examples of different accounts of the minimally sufficient compatibilist conditions for free will.
- 6 For examples interventionist accounts of causation, see Pearl (2009) and Woodward (2003).
- 7 Ismael (2013, 2016), Roskies (2012), and Campbell (2010).
- 8 Importantly, this soft-line response only works on manipulation arguments that feature an intentional manipulator (Deery and Nahmias 2017, p. 1273). If Diana thought she'd try her hand at manipulation and got lucky or if a natural event is responsible for Manny being the way he is (as discussed by Pereboom (2014)), then Deery and Nahmias must take the hard-line in response to such cases.
- 9 For another objection to Deery and Nahmias's soft-line response to the manipulation argument, see Tierney and Glick (2020).
- 10 For other interventionist approaches to responsibility that comes in degrees, see Chockler and Halpern (2004) and Halpern (2015, 2016). For non-interventionist causal theories of responsibility that come in degrees, see Beebe and Kaiserman (2020) and Kaiserman (2021).
- 11 In fact, she leaves open the possibility that freedom is grounded in a quasi-causal or other metaphysical relation that plays a similar role to that of causation (Sartorio 2016, p. 45).
- 12 See compatibilists such as Fischer and Ravizza (1998), McKenna (2008), and incompatibilists such as Pereboom (2001).
- 13 For another defense of the moral relevance of omissions, see Bernstein (2014, 2016).
- 14 For discussion of Frankfurt-cases featuring omissions, see: Swenson (2015, 2016), Fischer and Ravizza (1998), Clarke (1994, 2011, 2014).
- 15 In more recent work, Sartorio explores Frankfurt-style omission cases in much more detail (2017) and omissions more generally (2021, 2022).

- 16 Others come to different conclusions about proportionality when faced with competing accounts of causation. For example, Sartorio (2020), rather than embrace indeterminacy, rejects proportionality and argues that causal contribution does not come in degrees.
- 17 Some have begun to develop accounts of metaphysical relations distinct from causation, like production, that can help ground free will and moral responsibility. See Beckers and Vennekens (2018).
- 18 For a suggestion in a similar vein, see Kaiserman (2018).
- 19 I've reviewed several distinct conceptions of free will above. Perhaps some of these conceptions of free will can hang together in a single theory. For distinct conceptions of moral responsibility that hang together, see Watson (1996) and Shoemaker (2011).

Bibliography

- Beckers, S. and Vennekens, J. (2018). A principled approach to defining actual causation. *Sythese* 195: 835–862.
- Beebe, H. and Kaiserman, A. (2020). Causal contribution in war. *Journal of Applied Philosophy* 37 (3): 364–377.
- Bernstein, S. (2014). Omissions as possibilities. *Philosophical Studies* 167 (1): 1–23.
- Bernstein, S. (2016). Omission impossible. *Philosophical Studies* 173 (10): 2575–2589.
- Bernstein, S. (2017). Causal proportions and moral responsibility. In: *Oxford Studies in Agency and Responsibility* (ed. D. Shoemaker). 165–182, Oxford: Oxford University Press.
- Campbell, J. (2010). Control variables and mental causation. *Proceedings of the Aristotelian Society* 110: 15–30.
- Cartwright, N. (1979). Causal laws and effective strategies. *Noûs* 13 (4): 419–437. JSTOR.
- Chockler, H. and Halpern, J. (2004). Responsibility and blame: a structural-model approach. *Journal of Artificial Intelligence* 22: 93–115.
- Clarke, R. (1994). Ability and responsibility for omissions. *Philosophical Studies* 73 (2): 195–208.
- Clarke, R. (2011). Omissions, responsibility, and symmetry. *Philosophy and Phenomenological Research* 82: 594–624.
- Clarke, R. (2014). *Omissions: Agency, Metaphysics, and Responsibility*. Oxford: Oxford University Press.
- Deery, O. and Nahmias, E. (2017). Defeating manipulation arguments: interventionist causation and compatibilist sourcehood. *Philosophical Studies*, 174 (5): 1255–1276.
- Fischer, J.M. and Ravizza, M. (1998). *Responsibility and Control*. Cambridge: Cambridge University Press.
- Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy* 66 (23): 829–839.
- Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy* 68 (1): 5–20.
- Hall, N. (2004). Two concepts of causation. In: *Causation and Counterfactuals* (ed. J. Collins, N. Hall, and L.A. Paul). 225–276, Cambridge, MA: MIT Press.
- Halpern, J. (2015). Cause, responsibility and blame: a structural-model approach. *Law, Probability & Risk* 14 (2): 91–118.
- Halpern, J. (2016). *Actual Causality*. Cambridge, MA: MIT Press.
- Ismael, J.T. (2013). Causation, free will, and naturalism. In: *Scientific Metaphysics* (ed. H. Kincaid, J. Ladyman, and D. Ross), 208–235. New York: Oxford University Press.
- Ismael, J.T. (2016). *Why Physics Makes Us Free*. New York: Oxford University Press.
- Kaiserman, A. (2018). 'More of a cause': recent work on degrees of causation and responsibility. *Philosophy Compass* 13 (7): 1–10.
- Kaiserman, A. (2021). Reasons-sensitivity and degrees of free will. *Philosophy and Phenomenological Research*, 103: 687–709.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.
- McKenna, M. (2008). A hard-line reply to Pereboom's four-case argument. *Philosophy and Phenomenological Research* 77 (1): 142–159.

- Mele, A. (2013). Manipulation, moral responsibility, and bullet biting. *Journal of Ethics* 17 (3): 167–184.
- Nelkin, D. (2011). *Making Sense of Freedom and Responsibility*. New York: Oxford University Press.
- O'Connor, T. (1995). *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. New York: Oxford University Press.
- Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.
- Pereboom, D. (2001). *Living without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
- Roskies, A. (2012). Don't panic: self-authorship without obscure metaphysics. *Philosophical Perspectives* 26 (1): 323–342.
- Sartorio, C. (2016). *Causation & Free Will*. Oxford: Oxford University Press.
- Sartorio, C. (2017). The puzzle(s) of Frankfurt-style omission cases. In: *The Ethics and Law of Omissions* (ed. D. Nelkin and S. Rickless). 133–147, Oxford University Press.
- Sartorio, C. (2020). More of a cause? *Journal of Applied Philosophy* 37 (3): 346–363.
- Sartorio, C. (2021). Responsibility and the metaphysics of omissions. In: *Non-Being: New Essays on the Metaphysics of Nonexistence* (ed. S. Bernstein and T. Goldschmidt). 294–309, Oxford University Press.
- Sartorio, C. (2022). Juggling intuitions about causation and omissions. In: *Advances in Experimental Philosophy of Causation* (eds. P. Willemsen and A. Weigmann), 63–80, Bloomsbury Publishing.
- Shoemaker, D. (2011). Attributability, answerability, and accountability: toward a wider theory of moral responsibility. *Ethics* 121 (3): 602–632.
- Swenson, P. (2015). A challenge for Frankfurt-style compatibilists. *Philosophical Studies* 172 (5): 1279–1285.
- Swenson, P. (2016). The Frankfurt cases and responsibility for omissions. *The Philosophical Quarterly* 66 (264): 579–595.
- Tierney, H. and Glick, D. (2020). Desperately seeking sourcehood. *Philosophical Studies* 177: 953–970.
- Usher, M. (2020). Agency, teleological control and robust causation. *Philosophy and Phenomenological Research* 100 (2): 302–324.
- van Inwagen, P. (1983). *An Essay on Free Will*. New York: Oxford University Press.
- Vihvelin, K. (2013). *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. New York: Oxford University Press.
- Watson, G. (1996). Two faces of responsibility. *Philosophical Topics* 24 (2): 227–248.
- Wolf, S. (1990). *Freedom within Reason*. New York: Oxford University Press.
- Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Free Will and Reference*

SHAUN NICHOLS

Does free will exist? This is a candidate for the most interesting question we can ask about agency. The threat to free will derives from the idea that we are mistaken in how we think about our own agency. Typically, the mistake regarding free will is thought to be a false presupposition of libertarianism about our action (e.g., Strawson 1994). However, there is no direct line from “people’s notion of ϕ contains mistakes” to “ ϕ doesn’t exist.” Part of the issue, of course, is how significant the mistakes are. But even if we hold fixed the significance of the mistakes, this still might not settle the answer. The answer plausibly depends on subtle issues regarding the nature of reference. In this chapter, I’ll review how different views about reference can yield different conclusions about whether free will exists.

Of course many philosophers think that there is no mistake in the notion of free will. Some maintain that the notion of free will is thoroughly compatibilist and fits with the natural order; others maintain that the notion of free will is libertarian and libertarianism holds. For present purposes, I will set those happy-faced views aside and assume that there is some error in the descriptive content associated with FREE WILL. That is in keeping with the literature that I will discuss, but it excludes other important semantic approaches to free will that can’t be neatly classified this way (e.g., Turner 2013).

1 Reference and Arguments for Eliminativism

Free will is hardly the only philosophically interesting notion challenged by an eliminativist argument. We find something similar regarding the notions of morality (Mackie 1977; Blackburn 1985), race (Appiah 1995; Andreassen 2000), and belief (Stich 1983; Lycan 1988). As the philosophical discussion of these issues unfolded, a familiar pattern emerged. In each of these debates, eliminativists maintain that *K doesn’t exist* (where *K* might be morality, race, belief, etc.). Shortly after an eliminativist claim of this sort is made another group of philosophers adopt a *preservationist* position. In effect, they say, *Ks aren’t what we thought they were*. The question is then whether we should draw the eliminativist conclusion – *Ks doesn’t exist* – or the preservationist alternative – *Ks aren’t what we thought*. It is here that reference took on a central role.

The role of reference was particularly explicit in debates over eliminative materialism (e.g., Stich 1983, 1996; Lycan 1988; Bishop and Stich 1998), so let's start there. A central argument for eliminativism about belief promoted by Stephen Stich (among others) went roughly as follows:

- (1) "belief" is a term (or concept) in a folk theory;¹
- (2) that folk theory is massively mistaken;
- (C) therefore beliefs don't exist.

As presented, this argument is clearly not valid. The theory of reference might be brought in to supply some missing premises. In Stich's (1983) presentation of this kind of argument, he was quite clear in relying on a particular view about reference. He set out David Lewis's (1972) account of how theoretical terms get their reference (Stich 1983, pp. 17–21), and then uses this to push an eliminativist conclusion.

Lewis's account of the reference of theoretical terms is a *descriptivist* view of reference. According to early versions of descriptivism (Frege 1893; Russell 1919), the way terms refer to their objects is as follows:

D1: For a term *p*, *p* refers to *x* whenever users of *p* associate a certain description with the term, and *x* uniquely satisfies that description.

Lewis's (1972) descriptivist account of the reference of theoretical terms is more nuanced than D1; on Lewis's account, a term might refer when only some relevant *cluster* of the description is uniquely satisfied by *x*. But the view retains the crucial descriptivist element. If *nothing* satisfies a relevant cluster of the description associated with the term, then the term doesn't refer to anything. With this theory of reference in place, the above argument can be filled out. Stich argues that the associated description people have for "belief" (i.e., the folk psychological theory that gives reference to "belief") is deeply erroneous. Since the associated description is so badly mistaken, the term "belief" doesn't refer. When people use the term "belief," they aren't referring to anything. "Belief," like "witch," is a term that doesn't refer to anything. From there, it's plausible to conclude that beliefs, like witches, don't exist.² Lewis himself endorses a critical part of this argument, writing: "If the names of mental states are like theoretical terms, they name nothing unless the theory ... is more or less true" (Lewis 1972, p. 213). The key difference between Lewis (1972) and Stich (1983) concerns how much of the description associated with "belief" is false. Lewis is optimistic that the description is mostly right, whereas Stich has a more pessimistic view about the accuracy of the description, and this enables the eliminativist conclusion.

One response to eliminativism was to reject the pessimism about the folk concept of belief (e.g., Fodor 1987). However, another response was to reject the Lewisian account of how theoretical terms refer. We see this in an oft-cited passage from William Lycan. Lycan writes:

I am at pains to advocate a very liberal view ... I am entirely willing to give up fairly large chunks of our commonsensical or platitudinous theory of belief or of desire (or of almost anything else) and decide that we were just wrong about a lot of things, without drawing the inference that we are no longer talking about belief or desire. To put the matter crudely, I incline away from Lewis's Carnapian and/or Rylean cluster theory of the reference of theoretical terms, and toward Putnam's ... causal-historical theory. As in Putnam's examples of 'water,' 'tiger,' and so on, I think the ordinary word 'belief' (qua theoretical term of folk psychology) points dimly toward a natural

kind that we have not fully grasped and that only mature psychology will reveal. I expect that 'belief' will turn out to refer to some kind of information-bearing inner state of a sentient being, ... but the kind of state it refers to may have only a few of the properties usually attributed to beliefs by common sense. (1988, pp. 31–32).

Lycan promotes a causal-historical view of reference fixing. According to a classic version of the causal-historical theory, the way a term refers is as follows:

C1: For a term *p*, *p* refers to *x* whenever the use of *p* can be traced through a causal-historical chain of uses back to an initial baptism of *x*.

This theory of reference is permissive when it comes to descriptions. For on the causal-historical view, the term *p* might refer to *x* even if the description associated with *p* is rife with error and fails to pick out anything. What matters is that there is a chain of uses back to the entity or kind that was "baptized" when the term was introduced. As the term is transmitted from person to person, it continues to refer to the entity or kind that was baptized at the end of that causal-historical chain of transmission. As a result, people can have badly mistaken descriptions of the categories that their terms (or concepts) refer to.

The causal-historical theory of reference provides the most familiar example of a liberal view of reference fixing, i.e., one that is tolerant of lots of error in the description associated with a term. But it's important to note that descriptivist theories can vary on how liberal they are. An extremely restrictive descriptive theory might maintain that any associated errors nullify reference; a more liberal descriptive theory might maintain that even if there is very significant error, a term still refers so long as some subset of its associated description uniquely picks out a kind or category. The key distinction of interest is between theories of reference that are liberal (i.e., relatively tolerant of error) and those that are conservative (i.e., relatively intolerant of error). However, for ease of discussion, I will often focus on causal-historical as the key example of a liberal theory and a restrictive descriptivism as the key example of a conservative theory.

The eliminative materialist seems to embrace a conservative theory of reference, and says:

Beliefs don't exist. For *beliefs* are presupposed to be causally efficacious psychological states that are semantically evaluable; but there are no psychological states that have their causal efficacy *in virtue of* their semantic properties.

The preservationist replies:

Beliefs do exist. *Beliefs* are not as we supposed them to be – they are not causally efficacious *in virtue of* their semantic properties – but they are causally efficacious psychological states that are semantically evaluable.

The eliminativist maintains that the concept of belief carries a critical presupposition about semantic efficacy that is false; given a restrictive descriptivism, the fact that the concept carries this false presupposition might suffice to mean that beliefs don't exist. The preservationist maintains that the fact that the concept has this false presupposition isn't enough to drive the eliminativist conclusion; given a liberal theory of reference, there is reason to think that "belief" continues to refer despite the fact that it is associated with a critical presupposition that is false.

1.1 The Tree of Error

Error plays a key role in generating the threat of eliminativism. In philosophy, there is no shortage of such cases. Many of the most pressing questions in philosophy concern an eliminativist threat based on our false presuppositions surrounding a concept. And many philosophically interesting concepts – e.g., BELIEF, SELF, and FREE WILL – likely have significant error in their associated content.³ In the previous section, we saw how reference played a role in the debate over the existence of belief. But now I'd like to chart this territory more fully.

The typical eliminativist threat derives from the idea that a philosophically relevant concept has, as part of its associated content, an error. Perhaps the most familiar cases of such threats are ones in which the associated error is identified or specified. For instance, en route to his eliminativist conclusion, Mackie argues that people falsely presupposes that moral statements are objectively true. Here Mackie has specified what he takes to be an error at the heart of the folk concept of morality. Thus, this is a case of *specified error*. But sometimes we might hold that there is an error associated with a concept, without being able to identify what the error is. For example, the cognitive science of episodic memory is currently in flux, but it's likely that the right theory of episodic memory will be at odds with the folk notion of *remembering*. It's not exactly clear which parts of the folk view of memory are mistaken, but it's very likely that some parts of our view of memory are wrong. We can acknowledge that there are errors in the folk view of remembering, even though we can't yet identify them with confidence. Thus, when considering a concept, an error can be *specified* or *unspecified*, as depicted in the *Tree of Error* (see Figure 28.1).

The eliminativist/preservationist debate can play out on either branch of the Tree. An eliminativist can argue that the error associated with the concept yields a failure of reference, and hence that eliminativism follows for the kind under consideration. A preservationist, on the other hand, might maintain that reference succeeds despite the associated error, and hence no eliminativist conclusion follows. Most of the action in the free will literature concerns the *specified error* branch, but let's briefly consider the other branch. A familiar theme in philosophy of science has been that current scientific theories likely have a great deal of unspecified error. This assessment is based on our knowledge that past scientific theories were rife with error; intellectual modesty demands that we acknowledge that future scientists will

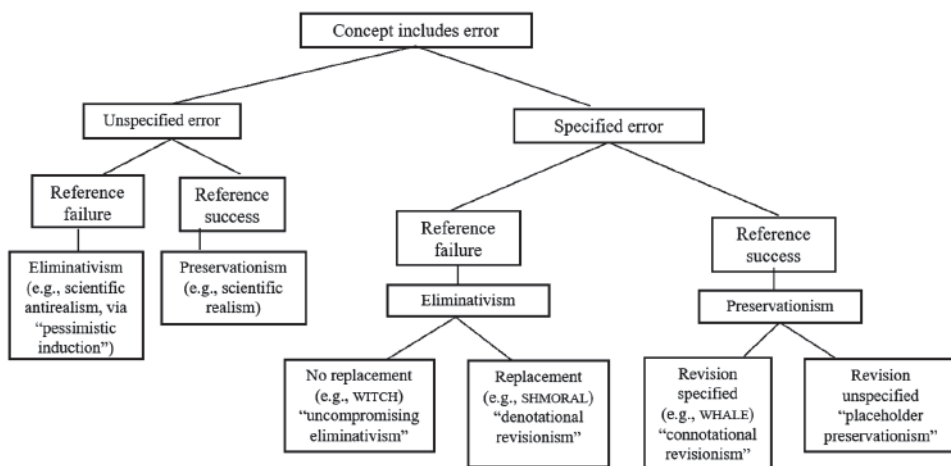


Figure 28.1 The tree of error.

regard our contemporary scientific theories as containing significant errors, even though we currently don't know what those errors are. Granting that there are unspecified errors is a critical step in arguments for scientific anti-realism based on the "pessimistic induction." Now, what about the role of reference? Reference failure is an explicit part of anti-realist arguments. Thus, Larry Laudan notes that prominent theories in the history of science are now taken to contain "central terms that (we now believe) were nonreferring" (Laudan 1984, p. 121). Stock examples include *phlogiston*, *caloric*, and *humors*. At this point, the anti-realist observes that our current theories likely contain significant errors (even though we can't currently specify them), and argues that we should expect the posits of current theory to go the way of *phlogiston* – the theoretical terms don't refer to anything and hence the posits don't exist. In the face of these arguments, scientific realists draw on liberal theories of reference to maintain that a scientific term can refer successfully even if the theory it is part of is full of error (Boyd 1983). This allows scientific realists to be preservationists about kinds in contemporary science, while still acknowledging that current theories are likely full of errors that we have yet to identify.

Turn now to the *specified error* branch of the Tree of Error. Again we get the division between preservationists and eliminativists. And again, the theory of reference plays a role in these arguments. Eliminativism is pressed by those who maintain that reference fails and preservationism is promoted by those who maintain that reference succeeds. But when the error is specified, we have a new set of options. If we know what the errors are, we are in a position to remedy the error. Or not. Let's take eliminativism first. In the sixteenth and seventeenth centuries, several authors promoted skepticism about the concept, WITCH. Ultimately, the view that prevailed held that witches don't exist because there is no one that has the occult powers and satanic contracts associated with witches. But no replacement was offered. We are well rid of that concept. We can call this "uncompromising eliminativism."

WITCH provides an example in which eliminativism prevailed and the concept was not replaced. But an eliminativist conclusion can be conjoined with a proposed replacement concept. For instance, Mackie argued for an eliminativist view of morality. He maintained that the concept of morality presupposes that morality is objective, and this false presupposition runs through all moral claims, yielding the result that all moral claims are false. One might concede Mackie's eliminativist conclusion and maintain that, unlike WITCH, we would be well served by having a replacement concept. Thus, we might maintain that even though the concept MORAL is empty, we can construct a related concept, SHMORAL, that has many of the social regulative features of morality but doesn't presuppose objectivity (see Blackburn 1985, p. 150). (See Figure 28.1.) Vargas calls this "denotational revisionism," since the replacement concept must have a different denotation than the empty concept. Vargas says that someone might "couple his eliminativism about morality (in the original sense) with a bit of referential re-anchoring, precisely because he thinks that the lurking-but-thus-far-not-usually-referred-to target property gets all the important things right" (2013, p. 88).

Now let's consider the other side. If reference of kind concept K is fixed a la the causal-historical theory (or some other liberal theory of reference), then K might refer despite the known errors associated with the concept. For a familiar example, take WHALE. The concept of WHALE was, for most of human history, associated with the mistaken view that the kind in question were a type of fish. In *Moby Dick*, Ishmael recounts and rejects Linnaeus's arguments for why whales are not fish. Instead, Ishmael says, "I take the good old fashioned ground that the whale is a fish, and call upon holy Jonah to back me." Despite the fictional narrator's loyalty to tradition, Linnaeus of course prevailed. But taxonomists did not draw the eliminativist conclusion that WHALE is an empty term and that whales don't exist.

Instead, we have a revised concept on which whales count as mammals, not fish. Since this revised concept still has the same denotation – it still picks out the same set of animals – this would count as “connotational revisionism” (Vargas 2013, p. 88)

Now let’s turn to the last leaf on the tree of error. One might know that concept K has an associated error, hold a liberal theory of reference that enables a preservationist view which affirms the existence of Ks, without proffering a revised version of the concept. That is, one might maintain that Ks exist, without trying to say what it is to be a K. We might call this *placeholder preservationism*. Such a preservationist move fits well with a certain cluster of views. If one has a causal-historical theory of reference, and one thinks of natural kinds as having essential natures, then one might maintain that while concept K is associated with error, it refers to a kind with an essential nature that we can’t currently articulate. This way of thinking about concepts converges with how cognitive psychologists describe lay categorization of natural kinds. Raccoons are taken to have some essence that is left unspecified – it’s marked by an “essence placeholder” (Medin and Ortony 1989) – but which distinguishes raccoons from skunks. In the above passage from Lycan, he seems to be taking such a placeholder preservationist view regarding *belief* when he writes, “I think the ordinary word ‘belief’ (qua theoretical term of folk psychology) points dimly toward a natural kind that we have not fully grasped and that only mature psychology will reveal” (1988, p. 32).

2 FREE WILL in the Tree of Errors

In debates over the existence of free will, few philosophers explore the “unspecified error” side of the tree. Among philosophers who maintain that there is error associated with the FREE WILL concept, the error is at least roughly specified – the error is typically taken to be a false presupposition of libertarian free will. Let’s consider how, given this specified error, positions regarding free will populate the tree of errors (see Figure 28.2).

Hard incompatibilists typically take an uncompromising eliminativist view about free will, maintaining that there is no such thing as free will and that we should not seek a replacement concept but face the truth (Strawson 1994; Pereboom 2001; Sommers 2012). An important element of this view, as was the case with WITCH, is that there are costly practices associated with the concept. People regarded as witches were persecuted and sometimes executed. The idea that people have free will is associated with holding them morally responsible and apt targets of punishment.

Vargas charts both denotational and connotational revisionism in the space of possibilities for free will (see also McCormick 2013). Vargas’s own revisionist proposal is neutral between the two. On Vargas’s revisionist notion of free will the critical feature is the capacity to respond to moral considerations.⁴ This revisionist account is intended to work regardless of what happens with the theory of reference. If the theory of reference allows for referential success, Vargas’s revisionist theory is an alteration of the original concept. If it turns out that the theory of reference yields eliminativism, then his revision is a replacement of the old FREE WILL concept; of course, even if one takes the replacement route, one can still use the term “free will” to pick out this replacement category.

Although denotational revisionism is an important leaf in the tree of error, few philosophers are explicitly committed to this view (as noted, Vargas allows for it but doesn’t commit). Connotational revisionism, by contrast, is familiar in the literature on free will and moral

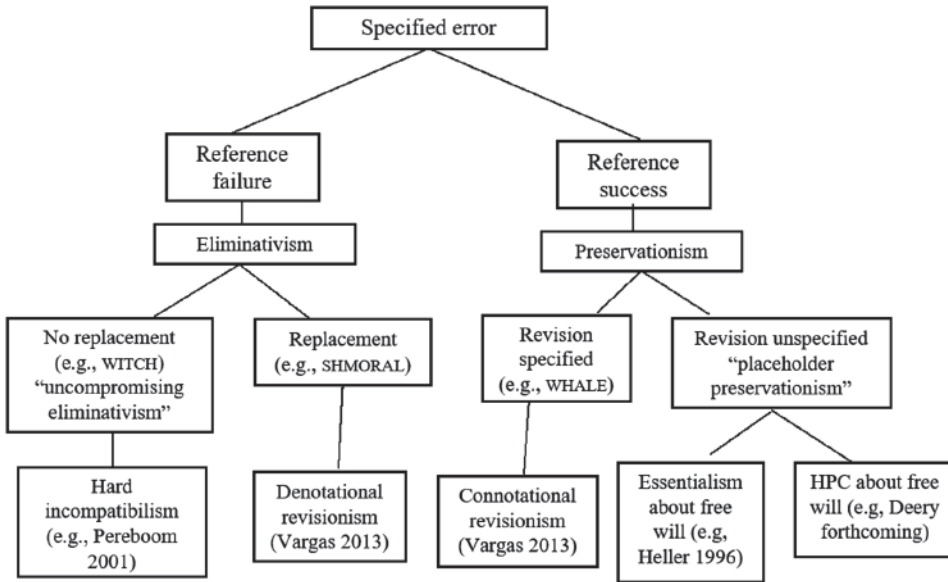


Figure 28.2 Free will in the tree of error.

responsibility. Frank Jackson, for instance, maintains that the ordinary notion of free will is incompatibilist and perhaps “embodies some kind of confusion” (2001a, p. 618); but Jackson claims that compatibilists nonetheless provide a successful analysis of the concept of free will. So, on his view, even though compatibilist accounts conflict with some folk intuitions, a compatibilist can give a proper analysis of the notion of free will (e.g., 2001b, p. 661) (see also Hurley 2000; McGeer 2013).

Finally, let’s turn to the last leaf: placeholder preservationism. Here one maintains that “free will” successfully refers without committing to exactly what free will is. This view is typically associated with rejecting a descriptivist theory of reference. As we saw, Lycan explicitly does this in his promotion of placeholder preservationism about belief. Mark Heller does something similar in his defense of compatibilism against counterexamples. Heller notes that on the Putnamian view, the description associated with a concept doesn’t determine the reference of the concept (333).⁵ One of Putnam’s familiar arguments here adverts to the possibility of robot cats. Heller channels Putnam:

To discover what it is to be a cat we put a cat before us and ask “what is essential to this thing?” We do not first have to have a concept of a cat or even know that the thing in question is called a “cat.” Discovering the essential nature of the kind becomes an empirical matter. If we discover that all the paradigm cats are robots, we are discovering something about cats, since the paradigm cases, and other creatures of the same kind, cannot fail to be cats. (Heller 1996, p. 334)

Heller goes on to argue that this Putnamian point has important lessons for “free will”:

By discovering that nothing fits our concept of free action, we would not have discovered that no actions are free. We would have discovered that free actions do not fit our concept of free action. This would be parallel to discovering that cats do not fit our concept of cat. (335).

Following Putnam's lead, Heller maintains that discovering the essential nature of a kind will be an empirical matter and this applies just as much for the kind *free action* as for the kind *cat* (334, 335).

The idea that kinds like *cat* have an essential nature is now quite controversial, at least among naturalistic philosophers (e.g., Hull 1965; Dupré 1981; Sterelny and Griffiths 1999), and the objections to kind essentialism plausibly extend to an essentialist account of *free action*. As a result, insofar as Heller's proposal depends on kind terms referring to essences, the view inherits considerable controversy (see also Daw and Alter 2001; Balaguer 2010). However, although Putnam links his anti-descriptivism about concepts with an essentialism about the kinds themselves, this link is not a necessary one. One can reject a descriptivist theory of reference without also holding that the kinds to which kind-terms refer have essential natures (see, e.g., Salmon 1979).

The fact that essentialism is not entailed by anti-descriptivism about reference opens up the possibility of a different kind of placeholder preservationism. Oisín Deery develops such an account (forthcoming). Instead of an essentialist account of kinds, Deery draws on the influential idea that kinds are homeostatic property clusters. Boyd characterizes this view of kinds as follows:

There is a family F of properties which are 'contingently clustered' in nature ... Their co-occurrence is the result of what may ... be described as a sort of *homeostasis*. Either the presence of some of the properties in F tends ... to favor the presence of others, or there are underlying mechanisms or processes which tend to maintain the presence of the properties in F, or both ... The clustering of the properties in F is causally important. (1988, p. 197)

Naturalistic philosophers take homeostatic property clusters to be an alternative to essences for a wide range of kinds, from species (Boyd 1991) to moral judgment (Kumar 2015). Deery applies the idea to free will, promoting a placeholder preservationist view that treats "free choice" as referring to a homeostatic cluster of features rather than an essence.

3 Settling the Debate

As we've seen, resolving the debate over whether free will exists plausibly depends partly on the theory of reference. If "free will" doesn't refer to anything, then free will doesn't exist. And different theories of reference seem to yield different verdicts on the question of whether "free will" refers to anything. The hard incompatibilist says "Free will doesn't exist"; the (connotational) revisionist says "Free will isn't what we thought it was." Who is right? The answer seems to depend partly on which theory of reference is correct. To settle the debate about whether eliminativism is right seems to require that we settle issues about the theory of reference. In this section, we'll explore some of the available options.

Eliminativist debates have been prominent historically, and in some cases the debates were resolved. It's worth reviewing how things played out. In the history of science, several chemists, including Stahl and Priestley characterized *phlogiston* as a substance released in combustion. This is reflected, for example, in Priestley's claim that if he had burned a pipe "in the open fire, the phlogiston would have escaped" (Priestley 1775, (vol. 2), p. 251). Lavoisier provided evidence for an alternative theory of combustion, on which no substance released in combustion. What happened next is well known, and it was not phlogiston preservationism. Instead, Lavoisier was an uncompromising eliminativist, saying that he aimed to "show

that Stahl's phlogiston is imaginary and its existence in the metals, sulfur, phosphorus and all combustible bodies a baseless supposition" (McKie 1936, p. 232). History has followed Lavoisier, and we deny the existence of phlogiston.

Something similar happened with the folk concept WITCH. The earliest witch skeptics in England were in fact preservationists (e.g. Scot 1584; Ady 1656). For instance, Reginald Scot, an important early skeptic, explicitly held a preservationist view: "My question is not (as manie fondlie suppose) whether there be witches or naie: but whether they can doo such miraculous works as are imputed unto them" (Scot 1584, p. 15). Within a century, skeptics took up an uncompromising eliminativist view (e.g., Wagstaffe 1671, pp. 83–84, Hutchinson 1718, p. 229). This eliminativist view has, of course, prevailed (Notestein 1911, p. 343).

Eliminativism about witches and phlogiston provides an important precedent. But they only tell part of the story. Other prominent episodes in the history of science illustrate a preservationist response. Consider the concept ATOM. Dalton, in keeping with earlier theorists and the very etymology of the word, held that atoms are indivisible. Rutherford discovered that Dalton's "atoms" were divisible. Divisibility is perhaps the most striking mark against Dalton's theory, but it was hardly the only problem. Indeed, Kitcher quips that "Almost everything Dalton maintained about atoms is wrong" (1993, p. 106). What happened on the heels of Rutherford's discovery? Obviously a preservationist view prevailed. Scientists continue to be committed to the existence of atoms, and we regard Dalton as just having been mistaken about the true nature of atoms.

When we turn to folk concepts, we can also find cases where preservationism prevailed. One case is especially interesting because it's closely related to WITCH – MAGICIAN. In the middle ages, *magicians* were regarded as individuals with supernatural powers (Thomas 1971, p. 41; see, e.g., King James I 1597, p. 31; Casaubon 1670, p. 111, Boulton 1715, p. 33).⁶ Yet we now use the concept MAGICIAN with no supernatural commitment. We just think that it's a mistake to regard magicians as having supernatural powers.

If we accept the authority of these historical cases, then it seems like sometimes when concepts are associated with error, eliminativism holds (e.g., witch and phlogiston); other times when concepts are associated with error, preservationism holds (e.g., magician and atom). What should we say about free will?

One somewhat radical possibility is pluralistic, building on the view that even for a given kind term, different views of reference can apply in different contexts. In recent experiments, Angel Pinillos, Ron Mallon, and I investigated the flexibility of people's reference-related judgments regarding kind terms (Nichols, Pinillos and Mallon 2016). We found that context made a difference to the reference-related judgment. In several experiments, we presented participants with information regarding a description of an animal called "catoblepas" drawn from medieval bestiaries. As we explained to participants, the concept of CATOBLEPAS was based on observations of wildebeests, but it was associated with an obviously empty description (e.g., catoblepas were supposed to have a death gaze). We found that people tended to disagree with the claim "Catoblepas exist" and agree with the claim "Catoblepas are wildebeests." This suggests that such statements trigger different reference conventions. For if participants were using the same reference convention for both statements, they would be saying something flatly contradictory – that catoblepas don't exist and they are wildebeests. Instead, it's plausible that when queried about "Catoblepas exist," people tend to disagree because they are adopting a conservative descriptivist way of thinking about reference, and the description is empty. By contrast, when queried about "Catoblepas are wildebeests," it's plausible that people are adopting a more liberal reference convention that tolerates the associated errors. People seem to fluidly switch reference conventions depending on the context. We think that

these results on lay intuitions suggest that our philosophical theory of reference should be pluralistic. Of course, one might maintain that ordinary people's reference-related judgments are irrelevant to the right theory of reference. But in that case, we need an account of what is relevant to determining the right theory of reference. That is a large issue. But it's safe to say that traditional theorizing about reference has relied primarily on commonsense reference-related judgments, and there is no alternative methodology that comes in as a close second.

If we incorporate a pluralist approach to reference into our interpretive practice, we can get a pacifistic answer to the metaphysical debate over the existence of free will (Nichols 2015, 2017). The eliminativist is assuming a conservative convention of reference and so, given that an important presupposition about free will is false, the eliminativist is right to say (under that convention for reference), "Free will doesn't exist." The preservationist on the other hand is adopting a more liberal convention of reference, and so the preservationist is right to say (under that convention for reference), "Free will isn't what we thought." Indeed, if we really do have these different reference conventions available, it seems most charitable to interpret eliminativists and preservationists accordingly. So, when Galen Strawson says "Free will doesn't exist," he is likely relying on a conservative convention for reference, and charity indicates that we should interpret him accordingly. Complementarily, when Bill Lycan says, "Free will is not quite what we thought it was," he is likely relying on a more liberal convention, and charity suggests that we should interpret him accordingly.

The foregoing pluralistic view about reference is not exactly widely adopted. There is a presumption that with "free will," as with any kind term, we should expect that a single view of reference applies, whether it be a liberal one or a conservative one (see, e.g., Vargas 2017; McCormick 2017). On this view, we might need to wait to find out the right theory of reference for "free will," but it will either be liberal or conservative, not both.

A rather different response maintains that even if one is pluralist about reference, eliminativism follows on both paths (Caruso 2015). Gregg Caruso agrees that if descriptivism is the right theory of reference, then "free will" doesn't refer because the description associated with "free will" contains a hopelessly false libertarian plank. Caruso maintains that the situation is also dire on the other path. He argues that even on a causal-historical approach to reference, the reference of "free will" is empty. For on a causal-historical approach, we need to identify the initial baptism that sets the reference of the term, and in the case of "free will," Caruso maintains, the initial baptism fixes on the phenomenology of free agency; this phenomenology itself, Caruso maintains, is an illusory libertarian phenomenology (2015, p. 2831).⁷ As a result, Caruso maintains that free will doesn't exist, no matter how the theory of reference goes.

4 Conclusion

In this chapter, I've canvassed the rough outlines of how the theory of reference might bear on debates about the existence of free will. The working assumption is that the lay notion of free will does contain some significant error, and the subsequent question is whether the term "free will" fails to refer in light of this error. As we've seen, different theories of reference yield different answers. However, it's important to note that many of these issues are only just now being explored by philosophers of free will. Hopefully the discussion of these matters will continue to flower.

Acknowledgment

I'd like to thank Joe Campbell, Oisín Deery, and Jason Turner for comments on an earlier draft of this. I'd also like to thank the John Templeton Foundation for supporting this project. Opinions expressed here are those of the author and do not necessarily reflect those of the Templeton Foundation.

Notes

- * Of course, this claim about the baptism is an empirical claim, and would require a good deal of scholarship to defend. Thanks to Joe Campbell for this point.
- 1 In the discussions below, I will assume that we can ask very similar questions about terms and concepts. In particular, I'll assume that the issues about empty reference apply both to concepts and to their corresponding linguistic expressions.
- 2 This last inference, from reference to metaphysics, has not gone unchallenged (see, e.g., Stich 1996; Mallon et al. 2009).
- 3 I will use the familiar notational expedient of using all caps for concepts.
- 4 "Free will is, roughly, the distinctive capacity of agents in virtue of which moralized praise and blame make sense. I argue that this capacity is best rendered in terms of the ability to recognize and respond to moral reasons" (Vargas 2011: 467).
- 5 Heller's actual terminology is a bit different from the terminology I've been using. First, on Heller's use of "concept," concepts include descriptive information. By contrast, I've followed the cognitive science literature and "concept" to mean a mental representation, without committing to the idea that the concept itself contains the description (which is denied by some philosophers of mind, e.g., Fodor 1998). So instead of speaking of concepts as containing descriptive information, I use the deliberately vague characterization on which the descriptive information is "associated" with the concept. Second, Heller characterizes the Putnamian view as denying that the concept determines "extension," whereas I've spoken in terms of reference. But extension is at least part of reference.
- 6 Some writers during this period did use "magicians" in roughly the modern sense (e.g., Wagstaffe 1671, p. 23; Webster 1677, pp. 166–167).
- 7 Of course, this claim about the baptism is an empirical claim, and would require a good deal of scholarship to defend. Thanks to Joe Campbell for this point.

Bibliography

- Ady, T. (1656). *A Candle in the Dark*. London: Robert Ibbitson.
- Andreasen, R.O. (2000). Race: biological reality or social construct? *Philosophy of Science* 67: S653–S666.
- Appiah, K.A. (1995). The uncompleted argument: Du Bois and the illusion of race. In: *Overcoming Racism and Sexism* (ed. L.A. Bell and D. Blumenfeld), 59–78. Lanham, MD: Rowman and Littlefield.
- Balaguer, M. (2010). *Free Will as an Open Scientific Problem*. Cambridge, MA: The MIT Press.
- Bishop, M.A. and Stich, S. (1998). The flight to reference, or how not to make progress in the philosophy of science. *Philosophy of Science* 65: 33–49.
- Blackburn, S. (1985). Errors and the Phenomenology of Value. In: *Morality and Objectivity*. (ed. Ted Honderich), 1–22. London: Routledge & Kegan Paul.
- Boulton, R. (1715). *Compleat History of Magick, Sorcery, and Witchcraft*. London, Great Britain: E. Curll, J. Pemberton, and W. Taylor.
- Boyd, R. (1983). On the current status of the issue of scientific realism. *Erkenntnis* 19: 45–90.
- Boyd, R. (1988). How to be a moral realist. In: *Essays on Moral Realism* (ed. G. Sayre-McCord), 181–228. Ithaca, NY: Cornell University Press.

- Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies* 61: 127–148.
- Caruso, G. (2015). Free will eliminativism: reference, error, and phenomenology. *Philosophical Studies* 172 (110): 2823–2833.
- Casaubon, M. (1670). Of credulity and incredulity in things divine and spiritual. London: S. Lownds.
- Daw, R. and Alter, T. (2001). Free acts and robot cats. *Philosophical Studies* 102 (3): 345–357.
- Deery, O. (forthcoming). Free choices, martians, and kinds.
- Dupré, J. (1981). Natural kinds and biological taxa. *Philosophical Review* 90: 66–90.
- Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.
- Fodor, J. (1998). *Concepts*. Oxford, UK: Oxford University Press.
- Frege, G. (1893). On sense and reference In: *Translations from the Philosophical Writings of Gottlob Frege* (ed. P. Geach and M. Black). Oxford: Blackwell (1952).
- Heller, M. (1996). The mad scientist meets the robot cats: Compatibilism, kinds, and counter examples. *Philosophy and Phenomenological Research* 56(2): 333–337.
- Hull, D.L. (1965). The effect of essentialism on taxonomy: two thousand years of stasis. *British Journal for the Philosophy of Science* 15: 314–326 & 16, 1–18.
- Hutchinson, F. (1718). *An Historical Essay concerning Witchcraft*. London: Knaplock and Midwinter.
- Hurley, S. (2000). Is responsibility essentially impossible? *Philosophical Studies* 99: 229–268.
- Jackson, F. (2001a). Précis of from metaphysics to ethics. *Philosophy and Phenomenological Research* 62 (3): 617–624.
- Jackson, F. (2001b). Responses. *Philosophy and Phenomenological Research* 62 (3): 653–664.
- James I, King of England. (1597). *Daemonology*. Edinburgh: Printed by Robert Walde-graue. Printer to the Kings Majestie. Cum Privilegio Regio.
- Kitcher, P. (1993). *The Advancement of Science: Science without Legend, Objectivity without Illusions*. New York, NY: Oxford University Press.
- Kumar, V. (2015). Moral judgment as a natural kind. *Philosophical Studies* 172 (11): 2887–2910.
- Laudan, L. (1984). *Science and Values*, Vol. 66. Berkeley, CA: University of California Press.
- Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy* 50 (3): 249–258.
- Lycan, W.G. (1988). *Judgement and Justification*. Cambridge, UK: Cambridge University Press.
- Mackie, J. (1977). *Ethics: Inventing Right and Wrong*. Penguin.
- Mallon, R., Machery, E., Nichols, S., and Stich, S. (2009). Against arguments from reference. *Philosophy and Phenomenological Research* 79 (2): 332–356.
- McCormick, K. (2013). Anchoring a revisionist account of moral responsibility. *Journal of Ethics and Social Philosophy* 7 (3): 1–19.
- McCormick, K. (2017). Why we should (n't) be discretionists about free will. *Philosophical Studies* 174 (10): 2489–2498.
- McGeer, V. (2013). Civilizing blame. In: *Blame: Its Nature and Norms* (ed. J.D. Coates and N.A. Tognazini), 162–188. Oxford, UK: Oxford University Press.
- McKie, D. (1936). *Antoine Lavoisier, the Father of Modern Chemistry*. Philadelphia PA: J.B. Lippincott.
- Medin, D.L. and Ortony, A. (1989). Psychological essentialism. In: *Similarity and Analogical Reasoning* (eds. S. Vosniadou, and A. Ortony), 179–195. Cambridge, UK: Cambridge University Press.
- Nichols, S. (2015). *Bound: Essays on Free Will and Responsibility*. USA: Oxford University Press.
- Nichols, S. (2017). Replies to Kane, McCormick, and Vargas. *Philosophical Studies* 174 (10): 2511–2523.
- Nichols, S., Pinillos, N.Á., and Mallon, R. (2016). Ambiguous reference. *Mind* 125 (497): 145–175.
- Notestein, W. (1911). *A History of Witchcraft in England from 1558 to 1718*. Washington: The American Historical Association.
- Pereboom, D. (2001). *Living without Free Will*. Cambridge: Cambridge University Press.
- Priestley, J. (1775). *Experiments and Observations on Different Kinds of Air*, vol. 2. London: J. Johnson.
- Russell, B. (1919). *Knowledge by Acquaintance and Knowledge by Description*: in *Mysticism and Logic*. London: George Allen and Unwin.

- Salmon, N.U. (1979). How not to derive essentialism from the theory of reference. *The Journal of Philosophy* 76 (12): 703–725.
- Scot, R. (1584). *The Discoverie of Witchcraft*. London: William Brome.
- Sommers, T. (2012). *Relative Justice: Cultural Diversity, Free Will, and Moral Responsibility*. Princeton, NJ: Princeton University Press.
- Sterelny, K. and Griffiths, P. (1999). *Sex and Death*. Chicago: University of Chicago Press.
- Stich, S.P. (1983). *From Folk Psychology to Cognitive Science: The Case against Belief*. Cambridge, MA: MIT Press.
- Stich, S.P. (1996). *Deconstructing the Mind*. New York, NY: Oxford University Press.
- Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies* 75 (1): 5–24.
- Thomas, K. (1971). *Religion and the Decline of Magic*. New York: MacMillan.
- Turner, J. (2013). (Metasemantically) securing free will. *Australasian Journal of Philosophy* 91 (2): 295–310.
- Vargas, M. (2011). Revisionist accounts of free will: origins, varieties, and challenges. In: *The Oxford Handbook of Free Will*, 2e (ed. R. Kane), 457–474. Oxford: Oxford University Press
- Vargas, M. (2013). *Building Better Beings*. Oxford, UK: Oxford University Press.
- Vargas, M. R. (2017). Contested terms and philosophical debates. *Philosophical Studies* 174 (10): 2499–2510.
- Wagstaffe, J. (1671). *The Question of Witchcraft Debated*, second edition. London: Edw. Millington, at the Pelican in Duck-Lane.
- Webster, J. (1677). *The Displaying of Supposed Witchcraft*. London: J.M.

Meaning in Life and Free Will Skepticism

DERK PEREBOOM

CORNELL UNIVERSITY

1 Skepticism about Free Will

We human beings are situated in a natural world of law-governed causes and effects. Our character and actions are conditioned by causes that we do not control, including our genetic make-up, our upbringing, and the physical environment. This much is evident. Some philosophers extend these considerations to an argument that all of our actions are conditioned by causes so as to rule out the possibility that we ever act freely. Still, we readily suppose that we do have free will. When we choose from among options for action, we typically assume that we do so with free will, and that when we choose, we instead could have chosen differently. The reasons for believing that these suppositions are false, and that no one has free will, are drawn from a number of sources. One traditional source is theological; everything that happens, including human action, is causally determined by the divine nature. Another is the naturalistic consideration that everything that happens, including all of our actions, is made inevitable by the remote past physical events together with the laws of nature.

In considering whether we have free will, we must keep in mind the term “free will” has a number of distinct senses, and the answer as to whether we have it may depend on which sense is in play. In a first sense, to have free will is to have alternative possibilities for choice and action:

Free will AP (for “alternative possibilities”): free will is an agent’s ability, at a given time, either to act or to refrain: that is, if an agent acts with free will, then she instead could have refrained at that time from acting as she did.

This is a classical sense of free will, and whether we have it, and whether it is possible both that we have **free will AP** and that causal determinism is true, is historically a core contested issue.¹

A second sense links free will to moral responsibility:

Free will MR (for “moral responsibility”): free will is an agent’s ability to exercise the control in acting required to be morally responsible for an action.

A Companion to Free Will, First Edition. Edited by Joseph Campbell, Kristin M. Mickelson, and V. Alan White.
© 2023 John Wiley & Sons, Inc. Published 2023 by John Wiley & Sons, Inc.

Many current participants in the free will debate contend that the sense that best serves to draw clear lines of difference is free will as the control in action required for our being morally responsible. More exactly, the participants disagree about whether it's possible that causal determinism is true and that we have the control in action required for being morally responsible for actions in a sense involving desert; specifically, in the *basic desert* sense. On the basic form of desert, an agent deserves the harm or pain of blame or punishment just because he has knowingly acted wrongly, and an agent deserves the benefit or pleasure of praise or reward just because she has acted in a morally exemplary way. The desert invoked here is basic because these claims about what agents deserve are fundamental in the sense that they are not justified by further considerations, such as the good consequences of implementing them (Feinberg 1970; Pereboom 2021, pp. 11–12; cf., 2001, p. xx, 2014, p. 2).

Assuming that “moral responsibility” refers to moral responsibility in the basic desert sense, the three traditional positions on the relation between free will and causal determinism are as follows:

Hard determinism: because determinism is true we cannot have the sort of free will required for moral responsibility.

(Soft determinist) Compatibilism: even though determinism is true, we can and do have the sort of free will required for moral responsibility.

Libertarianism: because determinism is false, we can and do have the sort of free will required for moral responsibility.

Hard determinists and libertarians contend that it's not possible that determinism is true and that we have free will.² Like the hard determinist, I agree that we don't have free will defined in this way.³ But this is not because I'm convinced of the truth of determinism, and that all of our actions are causally determined by preceding factors beyond our control, although I believe this may well be true. Crucially, in addition, I argue that any sort of indeterminism that has a good chance of being true is also incompatible with free will in the sense just defined. It's often argued that if our actions are not causally determined, they would then be random or chance events, and because we have no control over random or chance events, we can't be morally responsible for them. We'll soon examine an argument of this sort. Critics have voiced several practical concerns about the resulting skepticism about free will, that, for example, rejecting basic desert moral responsibility threatens various ways in which our lives can be meaningful. In what follows I address these objections. While some will remain unsatisfied, I believe that the responses are strong enough to make living without free will viable.

2 An Argument for Free Will Skepticism

I've developed a detailed argument for the skeptical position about free will (e.g., Pereboom 2001, 2014). Here I present the most recent version in outline. It features challenges to the competing positions, compatibilism and libertarianism, in turn. If these challenges are successful, only the skeptical view is left standing.

Compatibilists (of the sort at issue) maintain that even if all of our actions are causally determined by factors beyond our control, we can still be morally responsible in the basic desert sense for them. Compatibilists of this kind point out that causal determination is irrelevant to the commonplace criteria we use to ascertain whether people are blameworthy. In legal cases we may want to establish that the accused was not compelled by someone else to commit the crime, and that he was rational when he acted. Whether causal determinism is

true, these compatibilists contend, is not relevant to whether people are compelled or irrational, and moreover and more generally, whether determinism is true is never a consideration in court cases. Compatibilists set out conditions for moral responsibility that do not require the falsity of determinism, and they argue that satisfying such compatibilist conditions is sufficient for responsibility. Incompatibilists object that even if someone satisfies these compatibilists conditions, being causally determined by factors beyond her control rules out moral responsibility in the sense at issue. Does this amount to a standoff, or can progress be made in this discussion?

The best way to argue against the compatibilist option, in my view, starts with the intuition that in a case in which an agent is intentionally causally determined to act by, for example, neuroscientists who manipulate her brain by optogenetic stimulation,⁴ she will not be morally responsible for that action in the basic desert sense even if the compatibilist conditions are met. Next, we point out that there are no differences relevant to basic desert moral responsibility between this case and a possible case that features an agent who is causally determined to act in an ordinary naturalistic way. The conclusion is that an agent is not morally responsible in the basic desert sense if she is causally determined to act by factors beyond her control even if she satisfies the compatibilist conditions (Taylor 1962, pp. 45–46; Pereboom 1995, pp. 22–26, 2001, pp. 110–120, 2014, pp. 71–103; Kane 1996, pp. 65–69; Mele 2006, pp. 186–194).

In my multiple-case version of this argument, in each of four cases an agent commits a crime, murder, for self-interested reasons. All of the cases are designed so that the action conforms to the compatibilist conditions. For instance, the action meets a criterion advocated by David Hume (1739/1978): the agent is not compelled to act by other agents or constrained by other factors such as drugs. The action meets the rationality condition advocated by John Fischer (1994): the agent's desires can be modified by, and some of them arise from, his rational consideration of his reasons, and if he understood that the bad consequences for himself that would result from the crime would be much more severe than they are actually likely to be, he would have refrained from the crime for that reason. (cf., Wolf 1990, Wallace 1994, Nelkin 2011).

The manipulation cases serve to indicate that it's possible for an agent who knowingly acts wrongly to be morally non-responsible in the basic desert sense even if the compatibilist conditions are satisfied, and that, as a result, these conditions are insufficient for such moral responsibility, contrary to what the compatibilist claims. The argument gains force by setting out three such manipulation cases, each of which is progressively more like a fourth, in which the action is causally determined in an ordinary and natural way. The cases are designed with the aim that there be no difference relevant to basic desert moral responsibility between any two adjacent cases. So if it's agreed that the agent isn't morally responsible in the first case, this feature of the argument will make it difficult to affirm that he is responsible in the final, ordinary case:

Case 1: A team of neuroscientists possesses the technology and skill to manipulate Professor Plum's neural states remotely. In this particular case, they do so by pressing a button just before he begins to reason about his situation, which they know will result in a neural state that realizes a strongly egoistic reasoning process, which the neuroscientists know will deterministically result in his decision to kill White. Plum would not have killed White had the neuroscientists not intervened, because his reasoning would then not have been sufficiently egoistic to produce the decision to kill.

Case 2: Plum is just like an ordinary human being, except that a team of neuroscientists has programmed him at the beginning of his life so that his reasoning is often but not always

egoistic, and at times strongly so, with the intended consequence that in his current circumstances he will be causally determined to engage in the process of deliberation that results in his decision to kill White for egoistic reasons.

Case 3: Plum is an ordinary human being, except that the training practices of his community causally determined the nature of his deliberative reasoning processes so that they are often but not always egoistic. On this occasion, his deliberative process is exactly as it is in Cases 1 and 2: in his current circumstances he is causally determined to engage in the process of deliberation that results in his decision to kill White for egoistic reasons.

Case 4: All that occurs in our universe is causally determined by virtue of its past states together with the laws of nature. Plum is an ordinary human being, raised in normal circumstances, but his reasoning processes are frequently but not exclusively egoistic, and sometimes strongly so. In the current circumstances he is causally determined to engage in the process of deliberation that results in his decision to kill White, for egoistic reasons.

Case 1 features intentional, causally determining manipulation that is local in the sense that it takes place close to the time of the action. Of the four cases it is the one that most reliably elicits an immediate non-responsibility intuition. Case 2 is like Case 1, except that it restricts the deterministic manipulation to the beginning of Plum's life, and so with respect to Plum's action it is not local and more remote. Case 3 is distinctive in that the deterministic manipulation results from community upbringing. Case 4 is the ordinary deterministic case in which the causal determination of Plum's action is not intentional, but results from the past and the laws. Case 4 is a standard kind of case about which compatibilists claim that the agent is morally responsible despite being causally determined to act by factors beyond his control.

In Case 1 might Plum be basic-desert morally responsible for his action? It seems intuitive that Plum is a causally determined victim of conniving neuroscientists, and not responsible. Are there responsibility-relevant differences between Cases 1 and 2 that would justify the verdict that he is not responsible in Case 1 but is responsible in Case 2? As noted, it was my aim to construct the four cases so that it isn't possible to draw a difference relevant to the type of responsibility at issue between any two adjacent cases. Supposing this absence of relevant differences, if Plum is not responsible in Case 1, he isn't responsible in Cases 2, 3, and 4 either. I propose that the unifying best explanation for Plum's non-responsibility in the four cases is that in each he is causally determined to act by factors beyond his control. Hence the argument's anti-compatibilist conclusion.⁵

Next, the justification for the skeptical position about free will also requires arguing against the competing incompatibilist position, libertarianism. On libertarian views we do have the free will at issue, and required for an action's being freely willed is that it not be causally determined by factors beyond the agent's control. Here we'll examine the two most frequently endorsed kinds of libertarianism, *event-causal libertarianism* and *agent-causal libertarianism*.

On the event-causal libertarian position, actions are caused solely by *events*, conceived as substances having properties at times, such as *Abby wanting at noon today to give Rachel her medicine*. For an action to be freely willed, some type of indeterminacy in the production of the action by such events is a key requirement (Kane 1996; Ekstrom 2000; Balaguer 2010). The view that only events can be causes contrasts with one on which substances, such as atoms, machines, and agents can be causes, and not just events in which they have a part. Those who maintain that only events can be causes note that do speak as if substances can be causes, but maintain that when we clarify such speech, we will see that the event-causal view is right. For instance, imagine a car drives through a puddle of water and splashes you. We might say, "The car made you wet!" But more exactly, it's not the car that made you wet, but an event, the car's driving through the puddle at 725 5th Avenue in New York at 2:00 p.m. on the 7th

of April 2020 that had this effect. By contrast, according to *agent-causal libertarianism*, free will of the sort at issue is accounted for by agents who, specifically as substances, cause actions without being causally determined to do so. On this position, the causation that produces a free choice is not causation among events involving the agent, but is instead a case of the agent fundamentally as a substance causing a choice (e.g., O'Connor 2000; Clarke 2003).

A perennial objection to event-causal libertarianism is that if actions are undetermined in the way it proposes, agents will not have sufficient control in acting to secure moral responsibility for them. David Hume invokes this objection in his *Treatise of Human Nature* (1739/1978, pp. 411–412), arguing that if an action is not determined by some factor involving the agent, the action will not have the connection with the agent required for moral responsibility. I embellish this concern in the following way. To be morally responsible (in the basic desert sense) for an action, it's necessary that the agent have a certain robust kind of control in acting. Given the indeterminism specified by event-causal libertarian position, with the complete causal role of the causally relevant preceding events in place, it remains open whether action occurs. And furthermore, the part the agent plays in the production of the action is exhausted by these agent-involving events. Accordingly, nothing about the agent *settles* whether the action occurs, and this indicates that the agent lacks the control in acting required for the kind of moral responsibility at issue. Because the agent “disappears” at the critical point in the production of the action – at the point at which whether it occurs is to be settled – I call this the *disappearing agent argument* (Pereboom 2014, pp. 32–33, 2017).

Agent-causal libertarianism offers a remedy for this concern. If the agent herself, as a substance, is specified as a cause, she can have the role of settling whether the action occurs in such an indeterministic context. While on the event-causal alternative any causes involving the agent leave it open whether the action occurs, on the agent-causal view, the agent-as-substance can break the tie, and cause the action (or else refrain from doing so) without being causally determined to do so. More exactly, on agent-causal libertarianism agents possess a distinctive causal power, a power for an agent, fundamentally as a substance to cause an action without being causally determined to do so, and thereby to settle whether an action occurs (e.g., O'Connor 2000; Clarke 2003).

A frequently cited objection to agent-causal libertarianism is that it cannot be reconciled with scientific theory. Suppose that science reveals that the physical world is wholly governed by deterministic laws. Given this supposition, agent-causal libertarians may make room for their view by claiming, that we agents are non-physical beings, and that this allows us to have the agent causal power in an otherwise deterministic physical world. But this picture gives rise to a problem. On the path to a bodily movement resulting from an undetermined agent-caused decision, physical changes, for example in the agent's brain, occur. At this point we would expect to encounter divergences from the deterministic laws. Alterations in the brain that result from the causally undetermined decision would themselves not be causally determined, and would not be governed by deterministic laws. The agent-causalist may propose that the physical alterations that result from free decisions just happen to dovetail with what can be predicted on the basis of the deterministic laws, and then no event would occur that diverges from these laws. But this proposal features coincidences too wild to be credible. So it appears that agent-causal libertarianism cannot be reconciled with the physical world's being governed by deterministic laws. I argue that a very similar objection can be set out if the laws of physics are fundamentally probabilistic and not deterministic (Pereboom 1995, pp. 28–30, 2001, pp. 79–85, 2014, pp. 65–69).

The agent-causalist may now suggest that free decisions do in fact result in divergences from what we would expect given current theories of the physical laws. On this suggestion,

divergences from the deterministic or fundamentally probabilistic laws occur whenever we act freely. An objection to this proposal is that we currently have no evidence that such divergences actually occur. Accordingly, it appears that agent-causal libertarianism is not reconcilable with either deterministic or probabilistic laws of nature, and we have no evidence that divergences from these laws are to be found.

In summary, the manipulation argument undermines compatibilism; the disappearing agent argument dislodges event-causal libertarianism; and agent-causal libertarianism is not reconcilable with the laws of nature, while we have no evidence of events not governed by these laws. Free will skepticism is the position that remains standing. The objections most often raised for this position are practical: can we live with the belief that we lack the free will? The specific question I will now address is whether this view is reconcilable with meaning in life.

3 Deliberation, Decision, and Achievement

Given causal determination by the remote past and the laws of nature, the decisions we make as a result of deliberation would be inevitable due to causal factors beyond our control, and the effects of those decisions would also be inevitable by virtue of such factors. This in turn poses a threat to our sense of achievement, a core component of meaning in life. Do we ever really achieve anything if all of this is an inevitable result of factors beyond our control? (Sartre 1946/1956; Landau 2012).

However, there remains a sense in which how we deliberate and decide makes a difference to whether we achieve what we hope for. Imagine you were scheduled to interview for a new position five years ago, and you considered, due to time pressure, not preparing for it. But after deliberating – thinking of and weighing reasons for your options – you decided to prepare for the interview. Suppose that it was causally determined by the remote past and the laws of nature that you, upon deliberation, decided to prepare, and that as a result the interview went well, and you were offered the position, which you then accepted, and it has since been fulfilling and meaningful. It still may be true that if you hadn't prepared, the interview would have gone badly, and you wouldn't have been offered the position. In this sense your deliberation made a difference to how well you did in the interview and to whether you would receive offer, and to the resulting achievements and the meaning they provided for your life. This difference-making control contrasts with what Eddy Nahmias calls *bypassing* – that “our rational, conscious mental activity is *bypassed* in the process of our making decisions and coming to act” (Nahmias 2011, p. 556). As he points out, causal determination doesn't entail bypassing, because deliberation can be difference-making in the way I've just set out.

Still, on the supposition of naturalistic causal determination there are factors beyond our control – the remote past and the laws of nature – that made it inevitable how you deliberated, the decision to prepare that resulted, and the effects it had for the meaning of your life. This combination of claims may at first seem inconsistent, but it must be kept in mind that given causal determinism the difference deliberation can make is limited. Deliberation can indeed make a difference relative to other options for decision one was considering, but supposing naturalistic causal determinism, the non-occurrence of the options that weren't selected was still inevitable by virtue of factors beyond one's control.

Would our sense of achievement in our lives be undercut by these considerations? Suppose that you aim to succeed in your education or in your profession, and in fact you do very well. In such circumstances, we naturally have a sense of achievement, a sense that is a significant component for meaning in our lives. But now you come to believe that your success was an

inevitable consequence of the remote past and the laws of nature. Is your sense of achievement now been extinguished? I agree that it may seem diminished relative to a libertarian conception on which one's undetermined decisions are the ultimate origin of one's success. Still, as I've argued, supposing causal determination by factors beyond one's control, one's deliberation and decision may yet make a difference to one's success, and this can defensibly give rise to significant sense of achievement.

4 Moral Responsibility

One might think that a view of ourselves without moral responsibility for our actions detracts significantly from our conception of a meaningful life. However, in rejecting basic desert, the skeptic rejects only a resolutely backward-looking notion of moral responsibility. The skeptic may yet endorse a forward-looking notion of moral responsibility, and this indeed is what I have proposed (Pereboom 2014, 2017, 2021). On the version I endorse I develop and endorse, blaming is, in its paradigm cases, a kind of calling to account, and it is justified by forward-looking objectives, including the following:

The right of those wronged or threatened by wrongdoing to protect themselves and to be protected from immoral behavior and its consequences.

The good of reconciliation with the wrongdoer.

The good of the moral formation of the wrongdoer.

The retention of integrity of victims of wrongdoing.

Wrongful action is often harmful, and we have a right to protect ourselves and others from those who are apt to behave harmfully. Wrongdoing can also impair relationships, and we have a moral interest in reversing impairment by reconciliation. We value morally good character and action that results from it, and we accordingly have a stake in the moral formation of character when it is fraught by tendencies to misconduct. For those whose sense of integrity has been undermined due to being victims of wrongdoing, blaming can have a role in restoring that integrity.

There is an account of praise that parallels this notion of blame. Of the forward-looking aims just cited, the one most clearly amenable to praise is moral formation. We might praise an agent for a morally exemplary action to strengthen the disposition that produced it. Praise can also have a protective function; enhancing tendencies to behave morally by praising stands to reduce the incidence of harmful behavior. Corresponding to reconciliation is the celebration of success in a relationship, and praising may also have this role.

5 Relationships and the Reactive Attitudes

A further way in which our lives may acquire meaning is through fulfilling personal relationships. Love is a core emotion in such relationships. But it's been argued that love requires free will. To address this concern, we might first ask if for love to be valuable and provide meaning in our lives the person who is loved must have free will. Perhaps any valuable kind of love must be deserved by the beloved, and such desert must be grounded in freely willed decision and action. Against this, parents love their children independently of their having this sort of free will, and we think that such love is legitimate, valuable, and confers meaning (Pereboom 1995, p. 41, 2001, pp. 202–204, 2014, p. 190). Parents love their children when they are first

born, when they haven't had a chance to deserve love on the basis of free decision and action. Also, when adults love each other, it is frequently not grounded in what they've freely willed; appearance, intelligence, and affinities with persons or events in one's history may all have a part. At the same time, beneficial decisions and actions are often especially important for occasioning and maintaining love. But it would seem that love would not be threatened if we came to believe that such decisions and actions are never freely willed; they would be loveable whether or not the beloved is thought to deserve love for them (Pereboom 2001, pp. 188–198, 2014, pp. 190–193).

Perhaps for love to be valuable and to confer meaning in life, the love itself must be freely willed. If, for instance, love was intentionally causally determined through manipulation by another agent, one might think that it would have no value. John Milton, in his *Paradise Lost* (1665/2005), questions whether love for God would be valuable if it were intentionally causally determined by God. In Milton's poem, God gives a negative answer: "Of true allegiance, constant Faith or Love/Where only what they needs must do, appeared/Not what they would? what praise could they receive?/ What pleasure I from such obedience paid/When Will and Reason (Reason also is choice)/Useless and vain, of freedom both despoiled/Made passive both, had served necessity/Not me." Still, even if Milton is right about some such cases, it may be only the specific nature of the causal determination that is objectionable. Imagine that Rachel, using optogenetic neural stimulation, causally determines you to love her by making you inattentive to the ways in which she can be difficult in relationships. That would stand to be objectionable. However, suppose instead that you have a self-destructive tendency to love people who are apt to hurt you, and not to love those who would benefit you, partly because you have a tendency to overlook people's valuable characteristics. Suppose Xena slips a potion into your coffee that extinguishes this disposition, as a result of which you now appreciate her valuable characteristics and are causally determined to love her. How bad would that be? Perhaps what is unacceptable is not being causally determined to love by the other party *per se*, but rather how one is causally determined (Pereboom 2014, p. 192).

One might propose that free will has a key role in maintaining love over an extended period. Søren Kierkegaard (1843/1971) argues that marriage ideally involves a commitment to devote oneself to another that is continuously renewed, in the form of a decision that is repeated over time. A relationship with this feature is arguably desirable in virtue of being a voluntary expression of what one deeply cares about. We may now ask what of value would be lost if these repeated decisions were not freely willed? It would seem that here considerable value would be retained independently of free will. In sum, although it may initially appear that love that is freely willed is especially valuable, it is not clear exactly how free will would have a particularly desirable role in producing, maintaining, or enhancing love (Pereboom 2001, pp. 203–204, 2014, pp. 190–193).

P. F. Strawson (1962) sets out another way in which free will and the moral responsibility for which it is required may be essential to meaningful personal relationships. On his proposal, moral responsibility has its foundation in reactive attitudes such as resentment, indignation, guilt, and gratitude. Furthermore, given how we are psychologically constituted, the reactive attitudes and susceptibility to them are required for good personal relationships. Strawson suggests that the only alternative to relationships in which the reactive attitudes have a central role is a clinical and cold objectivity of attitude, which would seriously diminish the value of these relationships. If free will skeptics are right, however, the reactive attitudes are undercut because they presuppose a kind of moral responsibility we lack. Thus, given Strawson's view, free will skepticism would jeopardize the value of personal relationships and the meaning they might confer on our lives(cf. Wallace 1994, Shabo 2012). Strawson maintains that

our commitment to personal relationships rules out taking this skeptical threat seriously, and accordingly we have practical reason to ignore it.

I think that Strawson is right to claim that an objectivity of attitude of the sort he envisions would diminish the value of our relationships. However, I contend that he is mistaken to hold that such a stance would result if we were to eliminate or else disavow the central negative reactive attitudes, resentment, and indignation. These attitudes have a valuable role in personal relationships insofar as they serve to engage others when they act wrongly. But there are other attitudes available for such engagement that are not undercut by skepticism about free will. One such attitude is moral protest, a stance of opposition to an agent for having performed a specific immoral action or a number of immoral actions. Its function in our moral practice is primarily to engage wrongdoers by communicating opposition to wrongdoing of a general type, together with moral reasons for the agent to refrain from it (Pereboom 2021, pp. 27–53). Moral protest may be accompanied by other attitudes that don't presuppose deserved pain or harm, such as feeling hurt or disappointed about the wrongs others have done, and moral sadness or sorrow and concern for those who have acted wrongly (Pereboom 2001, pp. 199–202, 2014, pp. 178–186). I hold out for the possibility that this set of emotions serves to enhance personal relationships relative to the reactive attitudes that Strawson champions.

6 Meaning in a Universe without Free Will

One potential source of meaning in life is our conception of the universe and our place in it. The great theistic religions have provided views that potentially have this benefit, as have modern descendants of these religions whose notion of the divine is more impersonal. Such a view was endorsed by the Stoics, and is encountered in Judaism, Christianity, and Islam. Accepting a strong notion of divine providence involves the belief that everything that happens to us, to the last detail, accords with God's providential will. In the Stoic view, God causally determines everything that happens in accord with the good of the universe, although the nature of this good is imperfectly understood on our part. By identifying with this divine plan, we can be reconciled with the world's evils. Equanimity, and, more ambitiously, gratitude, replaces anger and despair as cosmic attitudes within our reach (Inwood 1985; Pereboom 1994, 2021; Bobzien 1998; Brennan 2005).

One might object that identification with the deterministic divine plan for the universe is too demanding for us given our limited capabilities. Suppose that one's role in the divine plan involves intense suffering up to the final end of one's life. In response, for many theists the divine plan aims not only at the good of the universe, but also at the good of specific individuals. As Marilyn Adams proposes, God is good to every person by ensuring each a life in which all suffering contributes to a great good within that very life (Adams 1999, p. 55). Then it might be, as Alvin Plantinga (2004) proposes, that God would know that if you were able to make the decision whether to accept the suffering of your life, and had sufficient knowledge of the divine plan, and had the right affections, you would accept that suffering.

These conceptions of a deterministic divine providential plan are nevertheless seriously challenged by the existence of evil, both moral and natural, that seems not to be in service of any good. If an all-powerful and strongly providential being exists, then that being would have the ability and the motivation not to preclude such evil. But there still seems to be such evil, so that conception faces a challenge. Responses to this problem of evil not decisive, but they arguably allow for rational hope that a strongly providential God exists (Jackson 2021; Pereboom 2021).

Is a hope of this general sort rational given a deterministic perspective that is not traditionally theistic or even atheistic? A conception that we see developed from the mid-eighteenth century onward focuses instead on the role of humanity. Humanity comes to be viewed as a force for inevitable and sustained progress in several dimensions. This theme was advanced by the German Idealists, of which Hegel is representative. Hegel (1824/1956) provides a vision of a deterministic process of human rational development over the course of history, recapturing the providential aspect of theological determinism by conceiving human development itself as divine.

Hegel's conception of the progress of human history has met with incredulous resistance due to the calamities of the first half of the twentieth century. Theodor Adorno's reaction "No universal history leads from savagery to humanitarianism, but there is one leading from the slingshot to the megaton bomb" (Adorno 1970, p. 312) is characteristic. At the same time, a theme that continues to find a place is human progress in history in the long term. John Dewey advocated a "common faith," a religious attitude broader than traditional theism, a "sense of nature as the whole of which we are parts, while it also recognizes that we are parts that are marked by intelligence and purpose, having the capacity to strive by their aid to bring conditions into greater consonance with what is humanly desirable" (Dewey 1934, p. 25). Let me note that Dewey's focus is broader than the human community, on "the enveloping world that the imagination feels is a universe" (Dewey 1934, p. 53).

Hope, by contrast with belief, or perhaps even faith, is arguably the appropriate attitude to invoke for humanity's survival in a thriving natural environment. We're not assured of this outcome, while it remains an open possibility. Hope is appropriate in contexts in which we are dependent for an outcome on factors beyond our control. As Katie Stockdale (2019) points out, hope reminds us of "the kinds of creatures we are: we are creatures who, because of the constraints we necessarily and contingently face as agents, must depend on factors external to ourselves for many of our desires to be fulfilled."

Theodore Parker, American transcendentalist minister, expresses a vision of continued human moral progress in his reflections on the abolition of slavery, using a now-familiar image:

We cannot understand the moral Universe. The arc is a long one, and our eyes reach but a little way; we cannot calculate the curve and complete the figure by the experience of sight; but we can divine it by conscience, and we surely know that it bends toward justice. Justice will not fail, though wickedness appears strong, and has on its side the armies and thrones of power, the riches and the glory of the world, and though poor men crouch down in despair, Justice will not fail and perish out from the world of men, nor will what is really wrong and contrary to God's real law of justice continually endure. (Parker 1853)

In his speeches, Martin Luther King also declared, quoting Parker, that "the arc of the moral universe is long, but it bends toward justice." Parker claims to have knowledge of long-term progress, but here hope is less demanding and more clearly justified.

What evidence do we have for Dewey's belief? In relatively recent history, humanity has faced extremely serious challenges and calamities: worldwide wars, mass genocide, terrorism, and pandemics. But we've confronted these challenges with impressive success from the perspective of humanity generally, although not from the point of view of the millions who suffered and the millions who perished. Many of the most oppressive regimes of the past century have vanished, medical science has made spectacular progress, technology has advanced very impressively, and knowledge in many fields has expanded massively. On balance, weren't the threats to humanity effectively confronted, and advance achieved? If so, we may rationally

hope is that our intelligence, skill, and moral resolve is sufficient for overcoming the challenges humanity will face in the future. And the rationality of this hope is consistent with causal determinism, and does not depend on our having free will (Pereboom 2021).⁶

Notes

- 1 I don't focus on **free will AP** in this entry, since it is less directly relevant to the issue of meaning in life than **free will MR**. Note that one might think that it's not possible that we have **free will MR** and that causal determinism is true, while it is possible that we have **free will AP** and that causal determinism is true. I suggest this position in Pereboom (2021), pp. 142ff. For an argument against the compossibility of causal determinism and **free will AP**, see van Inwagen (1983); for an argument for this compossibility, see Campbell (1997).
- 2 Kristin Mickelson (2015) distinguishes two distinct notions in the vicinity: on her proposal "incompossibilism" refers to the view that it's not metaphysically possible that causal determinism is true and that we have free will, while "incompatibilism" designates the more deeply committed explanatory view that causal determinism would explain why we lacked free will. I agree that this is an important distinction. But given that "compossibilism" is still an unfamiliar term, here I'll generally use "compatibilism" for what Mickelson calls "compossibilism."
- 3 Derk Pereboom (1995, 2001, 2014). Historical advocates of skepticism about free will include Spinoza (1677/1985), and Arthur Schopenhauer (1818/1961); and in recent decades, Galen Strawson (1986), Ted Honderich (1988), Bruce Waller (1990, 2011), Saul Smilansky (2000), Daniel Wegner (2002), Gideon Rosen (2004), Joshua Greene and Jonathan Cohen (2004), Shaun Nichols (2015), Thomas Nadelhoffer (2011), Benjamin Vilhauer (2012), Neil Levy (2011), Gregg Caruso (2012), Per-Erik Milam (2016), Stephen Morris (2018), and Farah Focquaert (2019).
- 4 <https://www.mightexbio.com/optogenetic-stimulation>. (Accessed on 20th May 2023).
- 5 If this is correct, it's not possible that causal determinism is true and that we have **free will MR**, and, in addition, causal determination explains why agents lack **free will MR**. Hence both incompossibilism and incompatibilism as Mickelson's (2015) defines them are true (see note 2). Objections to manipulation arguments are discussed in Pereboom (2014), pp. 71–103.
- 6 This contribution revises and updates Pereboom (2022).

Bibliography

- Adams, M.M. (1999). *Horrendous Evils and the Goodness of God*. Ithaca, NY: Cornell University Press.
- Adorno, T.W. (1970). *Negative Dialektik*. Frankfurt am Main: Suhrkamp.
- Balaguer, M. (2010). *Free Will as an Open Scientific Problem*. Cambridge: MIT Press.
- Bobzien, S. (1998). *Determinism and Freedom in Stoic Philosophy*. Oxford: Oxford University Press.
- Brennan, T. (2005). *The Stoic Life: Emotions, Duties, and Fate*. Oxford: Oxford University Press.
- Campbell, J.K. (1997). A compatibilist theory of alternative possibilities. *Philosophical Studies* 88: 319–330.
- Caruso, G.D. (2012). *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lanham, MD: Lexington Books.
- Clarke, R. (2003). *Libertarian Theories of Free Will*. New York: Oxford University Press.
- Dewey, J. (1934). *A Common Faith*. New Haven: Yale University Press.
- Ekstrom, L.W. (2000). *Free Will, A Philosophical Study*. Boulder: Westview.
- Feinberg, J. (1970). Justice and personal desert. In: J. Feinberg, *Doing and Deserving*. Princeton: Princeton University Press.
- Fischer, J.M. (1994). *The Metaphysics of Free Will*. Oxford: Blackwell Publishers.
- Focquaert, E. (2019). Neurobiology and crime: a neuro-ethical perspective. *Journal of Criminal Justice*, 65.

- Greene, J. and Cohen, J.D. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London, Series B-Biological Sciences* 359: 1775–1785.
- Hegel, G.W.F. (1824/1956). *The Philosophy of History* (trans. J. Sibree). New York: Dover Publications, 1956.
- Honderich, T. (1988). *A Theory of Determinism*. Oxford: Oxford University Press.
- Hume, D. (1739/1978). *A Treatise of Human Nature*. Oxford: Oxford University Press.
- van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Oxford University Press.
- Inwood, B. (1985). *Ethics and Human Action in Early Stoicism*. Oxford: Oxford University Press.
- Jackson, E. (2021). Belief, faith, and hope: on the rationality of long-term commitment. *Mind* 130: 35–57.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.
- Kierkegaard, S. (1843/1971). *Either/Or*, vol. 2 (trans. W. Lowrie). Princeton: Princeton University Press, 1971.
- Landau, I. (2012). Sartre's absolute freedom in being and nothingness. *Philosophy Today* 56 (4): 463–473.
- Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.
- Mele, A. (2006). *Free Will and Luck*. New York: Oxford University Press.
- Mickelson, K. (2015). A critique of Vihvelin's Threefold Classification. *Canadian Journal of Philosophy* 45 (1): 85–99.
- Milam, P.-E. (2016). Reactive attitudes and personal relationships. *Canadian Journal of Philosophy* 42: 102–122.
- Milton, J. (1665/2005). *Paradise Lost*. New York: W. W. Norton and Company.
- Morris, S. (2018). The implications of rejecting free will: an empirical analysis. *Philosophical Psychology* 31 (2): 299–321.
- Nadelhoffer, T. (2011). The threat of shrinking agency and free will disillusionism. In: *Conscious Will and Responsibility* (ed. L. Nadel and W. Sinnott-Armstrong), 173–188. Oxford: Oxford University Press.
- Nahmias, E. (2011). Intuitions about free will, determinism, and bypassing. In: *The Oxford Handbook of Free Will*, 2e (ed. R. Kane), 555–576. New York: Oxford University Press.
- Nelkin, D.K. (2011). *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.
- Nichols, S. (2015). *Bound*. New York: Oxford University Press.
- O'Connor, T. (2000). *Persons and Causes*. New York: Oxford University Press.
- Parker, T. (1853). Of justice and the conscience. In: *Ten Sermons of Religion by Theodore Parker* (ed. T. Parker). Boston: Crosby, Nichols and Company.
- Pereboom, D. (1994). Stoic psychotherapy in Descartes and Spinoza. *Faith and Philosophy* 11: 592–625.
- Pereboom, D. (1995). Determinism *Al Dente*. *Noûs* 29: 21–45.
- Pereboom, D. (2001). *Living without Free Will*. Cambridge: Cambridge University Press.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
- Pereboom, D. (2017). Responsibility, agency, and the disappearing agent objection. In: *Le Libre-Arbitre, approches contemporaines* (ed. J.-B. Guillon), 1–18. Paris: Collège de France.
- Pereboom, D. (2021). *Wrongdoing and the Moral Emotions*. Oxford: Oxford University Press.
- Pereboom, D. (2022). Hard determinism and meaning in life. In: *The Oxford Handbook of Meaning in Life* (ed. I. Landau). New York: Oxford University Press.
- Plantinga, A. (2004). Supralapsarianism, or 'O Felix Culpa'. In: *Christian Faith and the Problem of Evil* (ed. P. van Inwagen), 1–25. Grand Rapids: Eerdmans.
- Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives* 18 (Ethics): 295–313.
- Sartre, J.-P. (1946/1956). Existentialism is a humanism. In: *Existentialism from Dostoevsky to Sartre* (ed. W. Kauffman), 287–311. New York: Meridian Books, 1956.

- Schopenhauer, A. (1818/1961). *The World as Will and Idea* (later translated as *The World as Will and Representation*) (trans. R.B. Haldane and J. Kemp). Garden City, NY: Doubleday.
- Shabo, S. (2012). Where love and resentment meet: strawson's interpersonal defense of compatibilism. *The Philosophical Review* 121: 95–124.
- Smilansky, S. (2000). *Free Will and Illusion*. New York: Oxford University Press.
- Spinoza. (1677/1985). *Ethics* In: *The Collected Works of Spinoza* (ed. and trans. E. Curley), vol. 1. Princeton NJ: Princeton University Press.
- Stockdale, K. (2019). Social and political dimensions of hope. *The Journal of Social Philosophy* 50: 28–44.
- Strawson, G. (1986). *Freedom and Belief*. Oxford: Oxford University Press.
- Strawson, P.F. (1962). Freedom and resentment. *Proceedings of the British Academy* 48: 187–211.
- Taylor, R. (1962). *Metaphysics*. Englewood Cliffs: Prentice-Hall.
- Vilhauer, B. (2012). Taking free will skepticism seriously. *Philosophical Quarterly* 62: 833–852.
- Wallace, R.J. (1994). *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.
- Waller, B. (1990). *Freedom without Responsibility*. Philadelphia: Temple University Press.
- Waller, B. (2011). *Against Moral Responsibility*. Cambridge MA: MIT Press.
- Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.
- Wolf, S. (1990). *Freedom within Reason*. Oxford: Oxford University Press.

Free Will: Looking Ahead

ALFRED R. MELE

I was invited to write on “what lies ahead” in philosophical work on free will. If I were to set my target date ten years from now, a safe prediction might be increased attention to the topics that are getting the most attention now. Ten years is my target, and my own predictions will mainly be in this safe vein. But I want to do a little philosophy in this chapter too, as opposed to simply predicting what philosophy will do.

1 Skepticism about Free Will

When I raised my question about the future a few years ago on the Flickers of Freedom blog, many respondents predicted that skepticism about free will would increase markedly in the next few years. I am not confident that it will, but I do believe that philosophical attention to skepticism about free will is likely to increase significantly. The continued debate over whether such skepticism is justified promises to be lively and interesting.

Arguments for skepticism about the existence of free will divide into two broad kinds: theory-based and science-based. Familiar theory-based arguments are intended to show, wholly or primarily on theoretical grounds, either that some necessary condition for free will is unsatisfiable (without qualification or by human beings) or that some necessary condition for free will (probably) is never satisfied by human beings. Cases in point are versions of Galen Strawson’s “basic argument” (1986, 1994) that are explicitly about free will and a well-known argument by Derk Pereboom that features the idea that free will depends on indeterministic agent causation and such agent causation does not exist (2001, 2014).¹

Science-based arguments for the nonexistence of free will since the early 1980s often refer to work by neurobiologist Benjamin Libet, who makes the following much-discussed trio of claims: “The brain ‘decides’ to initiate or, at least, prepare to initiate [certain actions] before there is any reportable subjective awareness that such a decision has taken place” (1985, p. 536); “If the ‘act now’ process is initiated unconsciously, then conscious free will is not doing it” (2001, p. 62); “Our overall findings do suggest some fundamental characteristics of the simpler acts that may be applicable to all consciously intended acts and even to

responsibility and free will" (1985, p. 563). Associated with these claims is a skeptical argument that may be set out as follows:

- (1) In Libet-style experiments, all the decisions to act on which data are gathered are made unconsciously.
- (2) So probably all decisions to act are made unconsciously.
- (3) A decision is freely made only if it is consciously made.
- (4) So probably no decisions to act are freely made.²

Although this argument is specifically about decisions to act, it is easy to see how it might be extended to apply to all free actions.

Elsewhere I have argued that premises 1 and 2 are unwarranted (Mele 2009). I will not go into that again here, but I will mention an allegedly deeper point that is sometimes offered in response to my critique of those premises: namely, that Libet shows that decisions are produced by physical processes, and this itself shows that we never make decisions of our own free will. I will comment on the idea that the truth of substance dualism is required for the existence of free will after making and discussing some predictions.

Like theoretical arguments for the thesis that no human beings have free will, scientific arguments for that thesis purport to show that human beings do not satisfy some alleged necessary condition for free will. I predict that a significant part of increased philosophical attention to skepticism about free will will be focused on various alleged necessary conditions featured in the skeptical arguments. Does having free will depend on having indeterministic agent-causal powers (Pereboom 2001, 2014), on being "the cause of oneself ... in certain crucial mental respects" (Strawson 1994, p. 5), or on being or having an immaterial soul or mind? Questions of this kind will, I believe, receive increased attention (even if not all three of the questions I just asked do).

I also predict that this increased attention will come from two directions: traditional philosophy and experimental philosophy. Some philosophers use a combination of both approaches in seeking conceptual clarification. There will be more of that too, I believe.

Whether a given philosopher takes a purely traditional philosophical approach, an experimental approach, or a combined approach in investigating alleged necessary conditions for free will is explained largely by that philosopher's interests. A philosopher whose primary business is understanding "folk concepts" and who is interested in free will mainly as part of that project is likely to opt for a purely experimental approach. A philosopher who is fond of conceptual analysis and skeptical about the usefulness of lay judgments about free will is likely to opt for the traditional philosophical approach.

Philosophers may take the combined approach for various reasons. Compatibilists had heard for years that their view was out of touch with folk reality, and this claim was sometimes offered as some support for incompatibilism. This is part of what motivated some philosophers with a traditional interest in free will to get hard evidence about folk conceptions of free will (Nahmias et al. 2006). When they found what they regarded as evidence that most lay folk are compatibilists, they were in a position to use this finding at least to defend their compatibilist position against the charge that it was out of touch with lay thinking.

Another route to a combined approach is based on a philosopher's intended audience. Consider the following passage from neuroscientist P. Read Montague:

Free will is the idea that we make choices and have thoughts independent of anything remotely resembling a physical process. Free will is the close cousin to the idea of the soul – the concept that ‘you’, your thoughts and feelings, derive from an entity that is separate and distinct from the physical mechanisms that make up your body. From this perspective, your choices are not caused by physical events, but instead emerge wholly formed from somewhere indescribable and outside the purview of physical descriptions. This implies that free will cannot have evolved by natural selection, as that would place it directly in a stream of causally connected events. (2008, p. 584)

Montague is not alone among scientists in viewing free will as supernatural. Neuroscientist Michael Gazzaniga asserts that free will essentially involves a ghostly or nonphysical element and “some secret stuff that is YOU” (2011, p. 108). So it is no surprise that, in his view, “free will is a miscast concept, based on social and psychological beliefs ... that have not been borne out and/or are at odds with modern scientific knowledge about the nature of our universe” (219). (For other examples, see Mele 2012.)

Some philosophers with a traditional interest in free will also write for an audience that includes scientists who have published on free will. The number of such philosophers is growing. I predict it will grow significantly in the next ten years and that one effect will be increased attention to science in theoretical work on free will. If such a philosopher believes that free will is not supernatural and wishes to persuade Montague, Gazzaniga, and like-minded scientists of this, what approach is likely to be most productive? Using traditional philosophical methods, we have not been terribly successful at persuading one another to reject our compatibilism or our incompatibilism about free will. The same goes for our disagreements about whether having free will does or does not require having agent-causal powers, about whether compatibilists should agree or disagree that free action has an essential historical dimension, and so on. And it is a good bet that we tend to have more respect for our traditional methods than many scientists tend to do. When these observations are conjoined with the point that scientists respect good data, an experimental approach makes sense. Depending on one’s results, one may be in a position to show that those scientists who maintain that free will is supernatural have an idiosyncratic conception of free will. If this can be shown, these scientists can be moved to reflect on why they conceive of free will as they do. And when they look into that, they may find that the explanation does not run very deep and does not include anything that justifies the thesis that their conception of free will should be widely embraced.

A brief commentary on some experimental work is in order. Andrew Monroe and Bertram Malle conducted a study in which participants responded to the following request: “Please explain in a few lines what you think it means to have free will” (2010, p. 214). The 180 participants were undergraduates at the University of Oregon. Monroe and Malle report that “no assumptions of substance dualism ... were expressed” (216). Eddy Nahmias reports related findings:

a surprisingly low proportion of respondents: (1) *agreed* with the statement “Humans have free will only because they have nonphysical souls” (15–25%); (2) *agreed* with the statement “Our power of free will is something that is not part of our brain” (18%); or *disagreed* with the statement “It is because our minds are the products of our brains that we have free will” (only 13% when the statement followed a description of our brains as complex and unique, and still only

25% when the statement followed a description of the brain as mechanistic, governed by physical laws, and soon to be understood scientifically). (2011, n. 5)

I conducted some survey studies of my own on the present topic (Mele 2012). Here is one of the vignettes I used:

In 2019, scientists finally prove that everything in the universe is physical and that what we refer to as “minds” are actually brains at work. They also show exactly where decisions and intentions are found in the brain and how they are caused. Our decisions are brain processes, and our intentions are brain states. Also, our decisions and intentions are caused by other brain processes.

In 2009, John Jones saw a 20 dollar bill fall from the pocket of the person walking in front of him. He considered returning it to the person, who did not notice the bill fall; but he decided to keep it. Of course, given what scientists later discovered, John’s decision was a brain process and it was caused by other brain processes.

Almost three quarters (73.33%) of the participants said that John had free will when he made his decision, even though physicalism is very salient in the story. In this story, there is no place in the universe for nonphysical entities to be at work. The finding is at odds with the claim that ordinary usage of the expression “free will” treats free will as a supernatural power that depends on the existence of immaterial souls. My studies (Mele 2012), though they are not foolproof, definitely provide better evidence about ordinary usage of “free will” than does, for example, a randomly selected neuroscientist’s opinion about what that expression means.

Andrew Vonasch, Roy Baumeister, and I conducted a follow-up study as part of a much larger study (Vonasch, Baumeister, and Mele 2018). Participants read one of four randomly assigned probes. The two probes that are most pertinent for present purposes are reproduced below. (386 participants both completed the questionnaire and passed our comprehension check.)

Story A: Imagine that scientists finally discover that there are no souls. So, for example, they discover that you don’t have a soul, that your friends and neighbors don’t have souls, and so on.

Story B: Imagine that scientists finally discover that there are no souls. So, for example, they discover that you don’t have a soul, that your friends and neighbors don’t have souls, and so on. Like everyone else, John doesn’t have a soul. One day, he sees a twenty dollar bill fall out of the pocket of the person in front of him. He picks it up and keeps it for himself.

Participants responded to three statements about each story on scale from 1 to 7, where 1 is “strongly disagree” and 7 is “strongly agree.” The following statements accompanied Story A: People have free will in this situation; People can make conscious decisions in this situation; People are morally responsible for their actions in this situation. The statements accompanying Story B were specifically about John: John has free will in this situation; John can make conscious decisions in this situation; John is morally responsible for his actions in this situation.

Counting answers of 1, 2, and 3 as expressing disagreement and answers of 5, 6, and 7 as expressing agreement, we calculated the percentage of people who agreed with the statements and the percentage who disagreed. The results (rounded to the nearest whole number) are reported in the following table.

	Story A		Story B	
	Agree	Disagree	Agree	Disagree
Free Will	73%	18%	90%	4%
Conscious Decisions	75%	19%	91%	3%
Morally Responsible	70%	17%	86%	9%

Here again we have evidence that a substantial majority of nonspecialists do not believe that free will depends on souls, and the same goes for moral responsibility. More work on this topic is to be expected.

2 Manipulation Arguments and Original-Design Arguments

Stories featuring manipulation have generated considerable discussion in the literature on free will and moral responsibility. Sometimes they are used in arguments for the falsity of compatibilism about free will and determinism (Kane 1985, 1996; Pereboom 2001, 2014), and sometimes they are used in arguments for the falsity of a particular kind of compatibilist view (Mele 1995, 2006, 2019; Fischer and Ravizza 1998). In terms of possible worlds, compatibilism about free will and determinism, as I understand it (following standard practice), is the thesis that there are possible deterministic worlds in which free will exists. Incompatibilism is the denial of compatibilism. If free will is impossible, incompatibilism, so understood, is true. Incompatibilism also is true if free will is possible but absent in all deterministic worlds.³

Stories of the kind at issue are discussed more often in connection with moral responsibility than with free will. This is no accident, if intuitions about these stories tend to be stronger when moral responsibility is at issue. Given my assigned topic, I focus on free will here.

Compatibilists about free will disagree with one another about how agents' histories bear on a proper analysis of free action. The following passage expresses what has been called an *internalist*, *structuralist*, or *anti-historicist* view of the matter (notice that both moral responsibility and free action are mentioned here):

To the extent that a person identifies himself with the springs of his actions, he takes responsibility for those actions and acquires moral responsibility for them; moreover, the questions of how the actions and his identifications with their springs are caused are irrelevant to the questions of whether he performs the actions freely or is morally responsible for performing them. (Frankfurt 1988, p. 54)

To test claims of this kind, one can try to generate pairs of cases with the following features: each case highlights an agent who satisfies conditions the author claims to be sufficient for doing something freely, but, because of a difference in the two agents' histories, only in one case is the action at issue free. Regarding *any* proposed internalist sufficient conditions for free action, this testing strategy may be employed (perhaps successfully and perhaps not).

Some thought experiments feature manipulation-induced *radical reversals* in agents' hard-won values. Consider the following pair of stories.

Thoroughly Bad Chuck. Chuck enjoys killing people, and he "is wholeheartedly behind" his murderous desires, which are "well integrated into his general psychic condition" (Frankfurt 2002, p. 27). When he kills, he does so "because he wants to do it" (Frankfurt 2002, p. 27),

and “he identifies himself with the springs of his action” (Frankfurt 1988, p. 54). When he was much younger, Chuck enjoyed torturing animals, but he was not wholeheartedly behind this. These activities sometimes caused him to feel guilty, he experienced bouts of squeamishness, and he occasionally considered abandoning animal torture. However, Chuck valued being the sort of person who does as he pleases and who unambivalently rejects conventional morality as a system designed for and by weaklings. He freely set out to ensure that he would be wholeheartedly behind his torturing of animals and related activities, including his merciless bullying of vulnerable people, and he was morally responsible for so doing. One strand of his strategy was to perform cruel actions with increased frequency in order to harden himself against feelings of guilt and squeamishness and eventually to extinguish the source of those feelings. Chuck strove to ensure that his psyche left no room for mercy. His strategy worked. (See Mele 1995, pp. 162–163, 2006, p. 171.)

One Horrible Day. When Beth crawled into bed last night, she was one of the kindest, gentlest people on Earth. She was not always that way, however. When she was a teenager, Beth came to view herself, with some justification, as self-centered, petty, and somewhat cruel. She worked hard to improve her character, and she succeeded. When she dozed off, Beth’s character was such that intentionally doing anyone serious bodily harm definitely was not an option for her: her character – or collection of values – left no place for a desire to do such a thing to take root. Moreover, she was morally responsible, at least to a significant extent, for having the character she had. But Beth awakes with a desire to stalk and kill a neighbor, George. Although she had always found George unpleasant, she is very surprised by this desire. What happened is that, while Beth slept, a team of psychologists that had discovered the system of values that make Chuck tick implanted those values in Beth after erasing hers. They did this while leaving her memory intact, which helps account for her surprise. Beth reflects on her new desire. Among other things, she judges, rightly, that it is utterly in line with her system of values. She also judges that she finally sees the light about morality – that it is a system designed for and by weaklings. Upon reflection, Beth “has no reservations about” her desire to kill George and “is wholeheartedly behind it” (Frankfurt 2002, p. 27). Furthermore, the desire is “well integrated into [her] general psychic condition” (Frankfurt 2002, p. 27). Seeing nothing that she regards as a good reason to refrain from stalking and killing George, provided that she can get away with it, Beth devises a plan for killing him, and she executes it – and him – that afternoon. Her current view of things is utterly predictable, given the content of the values that ultimately ground her reflection. Beth “identifies [herself] with the springs of her action” (Frankfurt 1988, p. 54), and she kills George “because [she] wants to do it” (Frankfurt 2002, p. 27). If Beth was able to do otherwise in the circumstances than attempt to kill George only if she was able to show mercy, then, because her new system of values left no room for mercy, she was not able to do otherwise than attempt to kill George. When Beth falls asleep at the end of her horrible day, the manipulators undo everything they had done to her. When she awakes the next day, she is just as sweet as ever. (See Mele 2006, pp. 171–172, 2019, pp. 20–21.⁴)

Compare Beth’s killing of George with an arbitrarily selected premeditated killing by Chuck after he completes his heart-hardening project – his killing of Don.⁵ Each of these two killings satisfies conditions that Frankfurt deems sufficient for its being a free action.⁶ Some readers who find it plausible that Chuck kills Don freely and is morally responsible for killing him may find symmetrical judgments about Beth’s killing of George extremely implausible.

Not all internalist compatibilists endorse Frankfurt’s proposed set of sufficient conditions for free and morally responsible action. An alternative proposal may include the agent’s having been able at the time to do otherwise then or the agent’s being able at the time to

apprehend and act on moral reasons (or both). These abilities can be added to the two stories about killing.⁷

Consider a variant of the story about Chuck in which he has not quite completed his heart-hardening project (Mele 2019, pp. 24–25). He is considering killing Don just for the fun of it, but his internal condition, including his collection of values, is such that he is able to do otherwise in the circumstances than decide – and attempt – to kill him (at least in a respectable compatibilist sense of “able”) because he is able (in the same sense) to show mercy at the time, partly in response to an apprehension of some relevant moral reasons that have very little traction with him. Even so, Chuck decides to kill Don, and he succeeds.

In a companion story about Beth (see Mele 2019, p. 25), manipulators turn her overnight into a value twin of the present version of Chuck. She is considering killing George just for the fun of it. Her internal condition, including her collection of values, is such that she is able to do otherwise in the circumstances than decide – and attempt – to kill him (at least in a respectable compatibilist sense of “able”) because she is able (in the same sense) to show mercy at the time, partly in response to an apprehension of some relevant moral reasons that have very little traction with her. These abilities are not rooted in any preexisting values that survive the change in her. Instead, they are rooted in an implanted collection of values that matches the values in which Chuck’s parallel abilities are rooted. Beth decides to kill George, and she succeeds.

Here too, as in the earlier versions of these cases, some readers who find it plausible that Chuck kills Don freely and is morally responsible for killing him may find symmetrical judgments about Beth’s killing of George extremely implausible. If it were part of the story that in virtue of her continued possession of some of her earlier values, Beth was able to conquer her desire to kill George, intuitions might well shift. But, as I said, the pertinent abilities are not rooted in any surviving pre-transformation values of hers.

In the preceding stories featuring manipulation, a mature agent who has shaped his or her own character is targeted by manipulators. Here is a story of another kind that has been used in an argument for incompatibilism – an *original-design* story (from Mele 2006, p. 188).

Ernie’s Story. Diana creates a zygote *Z* in Mary. She combines *Z*’s atoms as she does because she wants a certain event *E* to occur thirty years later. From her knowledge of the state of the universe just prior to her creating *Z* and the laws of nature of her deterministic universe, she deduces that a zygote with precisely *Z*’s constitution located in Mary will develop into an ideally self-controlled agent who, in thirty years, will judge, on the basis of rational deliberation, that it is best to *A* and will *A* on the basis of that judgment, thereby bringing about *E*. If this agent, Ernie, has any unsheddable values at the time, they play no role in motivating his *A*-ing. Thirty years later, Ernie is a mentally healthy, ideally self-controlled person who regularly exercises his powers of self-control and has no relevant compelled or coercively produced attitudes. Furthermore, his beliefs are conducive to informed deliberation about all matters that concern him, and he is a reliable deliberator.⁸

Diana assembles *Z* as she does in Mary so that *E* will happen. Of course, in doing this, thereby ensuring that Ernie will bring about *E* by *A*-ing thirty years later, Diana ensures much more. A complete description of the state of the universe just after Diana creates *Z* – including *Z*’s constitution – together with a complete statement of the laws of nature entails a true statement of everything Ernie will ever do. In a version of this story, Diana does what she does in order to ensure everything that Ernie does (Mele 2006, pp. 189–190). Also, to block the claim that Ernie is not morally responsible for what he does only because another agent – Diana – is morally responsible for all of Ernie’s actions, it may be supposed (as in Mele 2006, p. 198, n. 6) that Diana is not morally responsible for anything because she is stark

raving mad and has no grasp of morality. These additional details should be regarded as official parts of Ernie's story.

Consider the following argument:⁹

- (1) Ernie never acts freely.
- (2) Concerning the freedom of the actions of the beings into whom the zygotes develop, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe.
- (3) So in no possible deterministic world in which a human being develops from a normal human zygote does that human being ever act freely.

Because the conclusion is specifically about human beings who develop from a normal human zygote, it does not take us all the way to incompatibilism. But, of course, typical compatibilists reject this conclusion.

We should not be surprised to find seasoned compatibilists rejecting premise 1. They might honestly report that they knew all along that they were committed to holding that an agent with Ernie's properties can act freely, and they might offer to explain why what they are committed to is true. In my own view (Mele 2006, pp. 192–193, 2019, pp. 84–85), this is the line compatibilists should take, and a powerful argument for the falsity of premise 1 would be especially welcome. Some philosophers have argued that a parallel version of premise 2 that is about moral responsibility rather than free will should be rejected (Waller 2014; Barnes 2015; Schlosser 2015; Deery and Nahmias 2017).

Ernie's story first appeared in print in 2006. It has generated a lot of discussion, and I predict that attention to the zygote-style of argument will not wane over the next ten years. I make the same prediction about manipulation arguments – especially those that target anti-historicist compatibilist views. I also predict that stories of the sort at issue will continue to be discussed more often in connection with moral responsibility than free will. But moral responsibility and free will are usually treated as importantly connected, and most philosophers who hold that Beth is not morally responsible for killing George would find it difficult to do so if they were to hold that she freely killed George.

3 Event-Causal Libertarianism

Libertarianism is the conjunction of two theses:

- (1) *The incompatibility thesis*: Free will is incompatible with determinism.
- (2) *The pro-free-will thesis*: There are actions that are or involve exercises of free will – *free actions*, for short.

Different libertarians take different positions on what free will requires beyond the falsity of determinism. *Agent-causal* libertarians contend that only beings with agent-causal powers can have free will. *Noncausal* libertarians argue that only uncaused actions can be (directly) free.¹⁰ *Event-causal* libertarians avoid appealing to agent causation, and they typically claim that paradigmatically free actions are indeterministically caused by their proximal causes.

I predict that attention to event-causal libertarianism will increase over the next ten years and that agent-causal and non-causal libertarianism will continue to receive about as much attention as they do now. The most fully developed event-causal libertarian view is found in

the work of Robert Kane (1996, 1999, 2014). Similar views (with some important differences) are advanced by Mark Balaguer (2010) and Christopher Franklin (2018) and floated by me in Mele 2006 and 2017.¹¹ (I say “floated” because I do not endorse libertarianism; I am officially agnostic about compatibilism.) Event-causal libertarianism is sometimes represented as an uninhabitable half-way house between compatibilism and agent-causal libertarianism. One thing that I believe will encourage some philosophers to jump in and develop event-causal libertarianism further is increased clarity about a feature of the existing arguments against it, namely, that they are seriously problematic. I cannot do justice to this claim in the space remaining, but some discussion is in order.

Pereboom claims that event-causal libertarianism fails because it “does not provide agents with any more control than compatibilism does” (2001, p. 56). And Randolph Clarke argues (RC1) that “the active control that is exercised on [an event-causal libertarian] view is just the same as that exercised on an event-causal compatibilist account,” adding (RC2) that the “view fails to secure the agent’s exercise of any further positive powers to causally influence which of the alternative courses of events that are open will become actual” (2003, p. 220).

We have in RC1 a partial basis for a control-featuring argument against event-causal libertarianism. I call it the *same-control argument*:

S1. Having free will depends on having a kind of active control that cannot be had in deterministic worlds and therefore cannot be captured by compatibilist accounts of free will.

RC1. “The active control that is exercised on [an event-causal libertarian] view is just the same as that exercised on an event-causal compatibilist account.”

So S3. Event-causal libertarianism is false.

One might take one’s lead in trying to ascertain whether RC1 is true from what John Fischer refers to as the distinction between “guidance control” and “regulative control” (1994, pp. 132–135). Guidance control can be exercised in deterministic worlds. An example is the control we normally exercise over how the cars we are driving move, if our world is deterministic. But regulative control, by definition, cannot be exercised in any deterministic world. When one *A*-s at *t*, one exercises regulative control over one’s *A*-ing only if there is another possible world with the same laws of nature and the same past up to *t* in which one does not *A* at *t*. Regulative control is not “just the same as” guidance control. And the active regulative control that is exercised on an event-causal libertarian view is not “just the same as” the active guidance control that is “exercised on an event-causal compatibilist account.” The former, by definition, is incompatible with determinism and the latter, by definition, is compatible with determinism. These facts preclude their being “just the same.” It can be said that event-causal regulative control is “just the same as” event-causal guidance control in a certain obvious respect – both are event-causal. But we need an argument to show us why we should believe that some true control requirement for free action demands something that no event-causal view of control can provide.¹²

Even though it is false that regulative control is “just the same” as guidance control, we still have the issue about “more control” that Pereboom raises (2001, p. 56).¹³ In discussions of comparative control in the free will literature, *direct control* is a prominent notion. Clarke writes: “Direct active control is exercised in acting, not before” (2003, p. 166). Timothy O’Connor reports that “exerting active power is intrinsically a direct exercise of control over one’s own behavior” (2000, p. 61). And Kane claims that agents exercise direct control over some of their choices (1996, p. 144). In these cases, Kane says, the agent’s exercise of control

is not “antecedent” to the choice; rather, it occurs “then and there,” when and where the choice is made.

The following argument features direct control. I dub it the *more-control argument*.

M1. Necessarily, if an agent’s world is deterministic, then even if he has as much control as an agent can possibly have in a deterministic world, he lacks free will.

M2. Necessarily, an agent with no agent-causal powers who has as much direct indeterministic control as can be had in the absence of agent-causal powers does not have a greater amount of control than an agent who has as much control as can be had in a deterministic world.

So *M3.* Necessarily, even an agent with as much direct indeterministic control as can be had in the absence of agent-causal powers lacks free will if he has no agent-causal powers.

If this argument is to be valid, it needs another premise. To see why, suppose that the following proposition is true:

P. An agent with direct indeterministic control and no agent-causal powers can have free will even if he does not have a greater amount of control than an agent who has as much control as can be had in a deterministic world.

There is no explicit contradiction in the conjunction of *M1*, *M2*, and *P*; but *P* entails that *M3* is false.

As far as I know, no one who argues that event-causal libertarianism is false on the basis of considerations of the sort featured in *M1* and *M2* has told us how to measure amounts of control or how to weigh deterministic and indeterministic control on the same scale.¹⁴ But suppose we grant *M2* for the sake of argument. If *M2* is true, the agents at issue might have the same amount of control or it may be that direct indeterministic control and any kind of deterministic control are incommensurable. In any case, in the absence of an argument against *P*, the more-control argument is at best incomplete.

Consider the claim that if two cars do not differ in horse power, the top speed of either cannot be greater than that of the other. This claim is false. Other features of cars are relevant to how fast they can move. Might amounts of control be like that in the sphere of free will? Someone might claim that if all relevant features of two agents that are not control features are equal, then if the agents do not differ in the amount of control they exercise at a time, either both act freely at that time or neither does. An argument for this claim may prove illuminating.

Premise *M1* of the more-control argument – the incompatibilist premise – is relevant in this connection. Someone who endorses it may say that agents in deterministic worlds do not exercise enough control when they act to act freely, but the same person may add that these agents do not exercise enough control because they do not have the right kind of control. What is the right kind? According to some philosophers, having the right kind of control requires having agent-causal powers – powers that they themselves are inclined to regard as impossible (Clarke 2003, p. 209) and powers the existence of which they say we have no evidence for (Clarke 2003, pp. 206–207) or weighty evidence against (Pereboom 2001, ch. 3, 2014, ch. 3).¹⁵ According to others, the right kind of control is a species of direct indeterministic control that is unsupplemented by any agent-causal powers (Kane 1996). An event-causal libertarian may claim that an agent can exercise enough of this kind of control

to act freely even if the amount of direct control he exercises does not surpass the greatest amount of control open to agents in deterministic worlds. (Instructions about how to weigh amounts of deterministic and indeterministic control on the same scale might prove useful for those who wish to assess this claim.¹⁶)

Another argument against event-causal libertarianism is Pereboom's "disappearing agent" argument, which he places in the family of "luck objections" (2014, p. 32). It runs as follows:

Consider a decision that occurs in a context in which the agent's moral motivations favor that decision, and her prudential motivations favor her refraining from making it, and the strengths of these motivations are in equipoise. On an event-causal libertarian picture, the relevant causal conditions antecedent to the decision, i.e., the occurrence of certain agent-involving events, do not settle whether the decision will occur, but only render the occurrence of the decision about 50% probable. In fact, because no occurrence of antecedent events settles whether the decision will occur, and only antecedent events are causally relevant, *nothing* settles whether the decision will occur. Thus it can't be that the agent or anything about the agent settles whether the decision will occur, and she therefore will lack the control required for basic desert moral responsibility for it. (32)

In lacking "the control" mentioned in the final sentence of this passage, the agent would lack control needed for deciding freely. The *main* problem I have with this argument is that nowhere in Pereboom 2014 is enough light shed on what it is to settle whether a decision will occur for readers to be confident that either of the following two claims is true: (1) an agent needs to be able to do this in order to decide freely and to have basic desert moral responsibility for his decision; (2) agents who lack agent-causal powers are unable to do this. (On this and other problems with the argument, see Palmer 2013; Mele 2017, ch. 8.)

Obviously, from the fact that leading arguments against event-causal libertarianism are problematic, it does not follow that event-causal libertarianism is true. My thought is that as philosophers become clearer on the problematic nature of these arguments, more of them will take an interest in developing an event-causal libertarian view that is superior to what already exists in the literature and to defending event-causal libertarianism against objections.

4 Conclusion

Projecting ten years into the future, I have made and discussed the following predictions about the philosophy of free will:

Skepticism: Attention to free will skepticism will increase. A significant part of this increased attention will be focused on various alleged necessary conditions featured in the skeptical arguments and will come from two directions: traditional philosophy and experimental philosophy.

Science: The number of philosophers with a traditional interest in free will who also write for an audience that includes scientists will grow significantly. One effect will be increased attention to science in theoretical work on free will.

Manipulation and original-design arguments: Attention to manipulation and zygote-style arguments will not wane, and the thought experiments will continue to be discussed more often in connection with moral responsibility than free will.

Libertarianism: Attention to event-causal libertarianism will increase significantly, and attention to alternative libertarian views will remain near current levels.

I close with a pair of predictions that I lack the space to develop:

Phenomenology: Interest in the phenomenology of free action will grow.

Moral responsibility: Increased attention will be paid to the idea that possible worlds with morally responsible agents do not need to include agents with free will.

I believe that even if all my predictions are false, free will will continue to be a lively topic of discussion for many years.¹⁷

Notes

- 1 The explicit skeptical conclusion of Strawson's basic argument is about true self-determination (1986, p. 29) or true moral responsibility (1994, pp. 5, 7), but he situates the argument in the literature on free will.
- 2 Libet maintains that although no decisions to *act* are freely made, decisions *not* to act on an intention to do something – decisions to veto intentions to act – may be freely made (2004, pp. 137–149). For discussion of Libet on vetoing, see Mele 2009, pp. 51–61 and 69–86.
- 3 There are also nonstandard definitions of “compatibilism” and “incompatibilism” (Levy 2011, pp. 1–2).
- 4 For a story about Chuck in which he is turned into a value twin of sweet Beth for a day, see Mele 2019, p. 29; for a similar story see Mele 1995, pp. 164–165.
- 5 Regarding the worry that Beth has been (temporarily) replaced by another person, see Mele 1995, p. 175, n. 22.
- 6 Frankfurt would not attempt to avoid this result by alleging insufficient psychic integration (see Frankfurt 2002, p. 28).
- 7 Some readers may view Beth and Chuck as being able to do otherwise than kill their victims – George and Don – even if they are unable to show mercy. Discussion of various compatibilist conceptions of ability is beyond the scope of this chapter.
- 8 Ernie is intended to meet conditions I have proposed as sufficient (but not necessary) for being morally responsible for performing a given action if compatibilism about determinism and moral responsibility is true (see Mele 2006, p. 188). Hence the reference to such things as ideal self-control and “unsheddable values” in this story. On these notions, see Mele 1995, pp. 116, 153–154.
- 9 This argument parallels an argument about moral responsibility set out in Mele 2019, pp. 101–102; see also Mele 2006, p. 189.
- 10 It is open to a noncausal libertarian to claim that some caused actions are indirectly free and inherit their freedom from the freedom of some uncaused free actions that are among their causes.
- 11 I have in mind the “daring libertarian” view that I float (Mele 2017, ch. 10). In Mele 2006, this is “daring soft libertarianism” (2006, ch. 5) minus only the “softness” – that is, minus only the willingness to entertain the idea that there is a kind of free action that is compatible with determinism.
- 12 On the same-control argument, see Mele 2017, pp. 142–147.
- 13 Clarke raises a closely related issue in RC2 above. I lack the space to discuss it here, but see Mele 2006, pp. 68–72 for a critique.

- 14 For a brief comparative discussion of amounts of indirect control in a particular connection, see Mele 2006, pp. 62–63.
- 15 Not all philosophers who claim that free will depends on agent causation are skeptics about agent causation. See O'Connor 2000.
- 16 On the more-control argument, see Mele 2017, pp. 143–146.
- 17 The chapter was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed here are my own and do not necessarily reflect the views of the John Templeton Foundation.

Bibliography

- Balaguer, M. (2010). *Free Will as an Open Scientific Problem*. Cambridge, MA: MIT Press.
- Barnes, E. (2015). Freedom, creativity, and manipulation. *Noûs* 49: 560–588.
- Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.
- Deery, O. and Nahmias, E. (2017). Defeating manipulation arguments: interventionist causation and compatibilist sourcehood. *Philosophical Studies* 174: 1255–1276.
- Fischer, J. (1994). *The Metaphysics of Free Will: An Essay on Control*. Malden, MA: Blackwell.
- Fischer, J. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.
- Frankfurt, H. (1988). *The Importance of What We Care About: Philosophical Essays*. New York: Cambridge University Press.
- Frankfurt, H. (2002). Reply to John Martin Fischer. In: *Contours of Agency: Essays on Themes from Harry Frankfurt* (ed. S. Buss and L. Overton), 27–31. Cambridge, MA: MIT Press.
- Franklin, C. (2018). *A Minimal Libertarianism Free Will and the Promise of Reduction*. New York: Oxford University Press.
- Gazzaniga, M. (2011). *Who's in Charge? Free Will and the Science of the Brain*. New York: Ecco.
- Kane, R. (1985). *Free Will and Values*. Albany: SUNY Press.
- Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.
- Kane, R. (1999). Responsibility, luck and chance: reflections on free will and indeterminism. *Journal of Philosophy* 96: 217–240.
- Kane, R. (2014). New arguments in debates on libertarian free will: responses to contributors. In: *Libertarian Free Will* (ed. D. Palmer), 179–214. Oxford: Oxford University Press.
- Levy, N. (2011). *Hard Luck*. Oxford: Oxford University Press.
- Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8: 529–566.
- Libet, B. (2001). Consciousness, free action and the brain. *Journal of Consciousness Studies* 8: 59–65.
- Libet, B. (2004). *Mind Time*. Cambridge, MA: Harvard University Press.
- Mele, A. (1995). *Autonomous Agents: From Self-Control to Autonomy*. New York: Oxford University Press.
- Mele, A. (2006). *Free Will and Luck*. New York: Oxford University Press.
- Mele, A. (2009). *Effective Intentions: The Power of Conscious Will*. New York: Oxford University Press.
- Mele, A. (2012). Another scientific threat to free will? *Monist* 95: 422–440.
- Mele, A. (2017). *Aspects of Agency: Decisions, Abilities, Explanations, and Free Will*. New York: Oxford University Press.
- Mele, A. (2019). *Manipulated Agents: A Window to Moral Responsibility*. New York: Oxford University Press.
- Monroe, A. and Malle, B. (2010). From uncaused will to conscious choice: the need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology* 1: 211–224.
- Montague, P.R. (2008). Free will. *Current Biology* 18: R584–5.
- Nahmias, E. (2011). Why 'Willusionism' leads to 'bad results': comments on Baumeister, Crescioni, and Alquist. *Neuroethics* 4: 17–24.
- Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research* 73: 28–53.

- O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.
- Palmer, D. (2013). Event-causal libertarianism: two objections reconsidered. In: *Free Will and Moral Responsibility* (ed. I. Haji and J. Caouette), 98–122. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Pereboom, D. (2001). *Living without Free Will*. New York: Cambridge University Press.
- Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press.
- Schlosser, M. (2015). Manipulation and the zygote argument: another reply. *Journal of Ethics* 19: 73–84.
- Strawson, G. (1986). *Freedom and Belief*. Oxford: Clarendon Press.
- Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies* 75: 5–24.
- Vonasch, A., Baumeister, R., and Mele, A. (2018). Ordinary people think free will is a lack of constraint, not the presence of a soul. *Consciousness and Cognition* 60: 133–151.
- Waller, R. (2014). The threat of effective intentions to moral responsibility in the zygote argument. *Philosophia* 42: 209–222.

Epilogue: Free Will Zombies

V. ALAN WHITE

1 A Parable

Libertarians were understandably jubilant when the invention of the I-detector was announced. This device detects in people the presence or absence of an indeterministic process that – as it turned out – always and only attaches to ordinary capacities for reason in any exercise of choice, and thus “effectively establishes a powerful correlation indicating the presence or absence of libertarian free will” (so said the Surgeon General of Bioethics on a label eventually affixed to the device). In other words when the detector shows that indeterminism is absent by producing no reading, effective reasoning capacity is absent too (as empirically demonstrated), so it maps the familiar territory of the presence or absence of basic-desert-bearing responsibility in libertarian terms.¹ Compatibilists in particular grumbled that this demonstrated strict correlation of indeterminism with reasoned choice did not yield a plausible mechanism for how it works, but libertarians just snapped back that Humean causation and laws hadn’t troubled them much before – why should similar rigid correlations be worrisome now? Admittedly there were still many conferences headlining the back-and-forth between event-causal libertarians and their agent-causal cohorts because the I-detector data was epistemically compatible with both, but not many conferences were funded for hard incompatibilists or compatibilists or free will skeptics, though they still gathered together often for informal symposia (in the Platonic sense).

One product of this discovery was a re-energized emphasis on retribution in both personal and public policy matters. Critics complained that even the granted truth of libertarianism was conceptually insufficient to justify retributive practices, but advocates shoved I-data back in their faces as a plausible basis for a claim to moral ultimacy as a necessary condition for separating moral wheat from chaff.²

So things continued in this way, and the I-detector proved increasingly indispensable for matters of justice. Then something unexpected happened: evolution threw a wrench in the gears with a mutation. Over time a population appeared who were outwardly and (as far

as self-reports can tell us anything) inwardly just like people with libertarian free will, but the I-detector yielded null results when trained on them. Of course initially these individuals were prejudicially thought to lack ordinary capacities for reasoning and treated accordingly. However, it eventually became apparent that that was not the case: double-blind studies proved that these individuals were indistinguishable in reasoning capacities from those with libertarian free will.³ They pondered choices in the same kinds of way, sometimes choosing well, sometimes ill and in comparable proportion to libertarians.⁴ Naturally a derisive term emerged to describe them. They were “free will zombies.”

Of course the I-detector became useless when attempting to assign libertarian-style ultimate responsibility in any practical sense to free will zombies, especially when it became a tactic for some more unsavory zombies to feign insanity in order to escape retributive blame with the assist of the I-detector.⁵ But this was also something of a philosophical windfall for those few struggling compatibilists and hard incompatibilists and free will skeptics left (usually employed as baristas rather than academics since such non-zombie “free will flat-earthers” chose to deny obvious truth ironically of their own free will⁶). They questioned whether in cases where libertarians and zombies behaved identically by all standards of assessment except for different I-detector readings, would it be the case that retributive treatment (based primarily on incompatibilist grounds) is fairly applied to zombies, since they lack what had become an accepted feature of praise- and blameworthiness? Would two systems of justice (perhaps modeled on the juvenile/adult systems) be necessary for equivalently rational defendants who were precisely distinguishable in just one way – the presence or absence of indeterministic I-properties? Or, were distinguishable zombies *really people* worthy of retributive concern? And perhaps more pertinently some whispers grew into a shout: was the system of retributive justice based on libertarianism somehow faulty from the get-go?

2 The Parable’s Reflections in Our Mirrors

Typically discussions of free will pose distinct accounts of human nature as inhabiting very different possible worlds – deterministic, indeterministic, fatalistic, mystertian, conceptually skeptical, and so on. Various sets of intuitions and argumentative schema align with one world as favored and others thus decried, and dialectical stalemate among disputants is a familiar ellipsis.⁷ What gets lost in the fray is the prospect of finding some direct way to compare such accounts in more *practical* terms. My little story about zombies attempts to remedy that lack of comparison, at least in part.

Thus libertarianism is planted squarely in just one world and given means (the I-detector) to rationally justify – at least initially by strict empirical correlation ala smoking and cancer – belief in its truth. Consequently, retributive practices based on libertarian intuitions are justified to that extent as well. But then a minority population of zombies come on the scene, who appear to be as much human as anyone else and in particular are reasonable beings indistinguishable from libertarians save having no discernible indeterministic processes involved in their rational choices. One might assume that this indicates that zombies are deterministic, but it is sufficient to say that whatever they are, they are not I-type libertarian beings.

So when we have a badly behaving zombie on our hands, what is the most rational way to deal with that? Since zombies at best are only capable of having some form of compatibilist free will, or at worst none at all, then libertarian-based retributive practices appear to be off the table. Of course any forward-looking consequentialist-based deterrence or treatment used for libertarian criminals would be extended to criminal zombies on presumably equally

just grounds, but not libertarian retribution. Still, if such retribution continued to be applied to I-verified libertarian criminals, then a two-tiered justice system would exist.

In that case, no doubt some libertarians (probably defense attorneys) would claim reverse-discrimination. After all, did they (or their libertarian clients) ask to be endowed with I-properties? Why should they, even as criminals, be treated differently – and frequently more harshly – for something they could not help but possess but otherwise made them no different from better-treated zombies? Of course they seldom stressed that praise heaped on them for trivial acts of kindness *might* mean so much more than that accorded Nobel Prize-winning zombies, but praise after all is so eminently ignorable compared with questions of doing hard time and capital punishment.⁸ Note, however, that such protests would be made not primarily on metaphysical grounds, but *practical* ones, because libertarians would thus claim that their I-properties were regarded as something meaningful for justice when it was really only an irrelevant point of reverse prejudice as compared with the more humane treatment of zombies, with whom they shared so much otherwise.⁹

Other libertarians of course might wallow in that very difference and make I-properties a point of pride. “Take your stinking paws off me you damn dirty zombie!” Nevertheless, it seems mind-boggling that one such small difference could legitimize I-elitism, especially since there had never been a conclusive account of *exactly how* – and now more strongly *if* – I-properties were significantly involved in making rational decisions.¹⁰

I should hope that soon cooler heads would prevail, and everyone would realize that for all *practical* purposes the minor difference of those I-properties was irrelevant for being considered human and fully responsible, thus resulting in something like real moral progress.¹¹ If free will zombies really possess all other salient features for full inclusion in the moral sphere, then it is mere prejudice to call them out as “zombies” for differential – and what reverse-discrimination-zombie-envying libertarians would even call *preferential* – moral treatment. Zombies are rational decision-makers who deserve attributions of responsibility, even if less pointedly than in the sense of libertarian basic desert *ultimacy* riveted to I-properties, which palpably the existence of zombies demonstrates is only *circumstantially* allied to reason in I-verified instances. This makes justification for retributive responsibility *practically* incoherent, and favors by default consequentialist or compatibilist theories of responsibility and punishment for everyone based on the preeminent zombie example.¹²

What is the moral of this tale? Just this: the free will problem has too long ignored the role that *moral pragmatism* might play in resolving it¹³, especially given that we in our actual world are in the same kind of metaphysical epistemic depravity about the nature of free will reflective of that in my parable.¹⁴ Once brought squarely into the picture as with this one-world direct comparison of libertarians with free will zombies, we clearly see that it makes no practical moral difference whether we in our own world have I-properties or turn out to be just those I-less zombies, and thus emphasis on such a property that is taken by itself to entail retributivism is unjustified.¹⁵

Notes

- 1 Throughout the rest of this paper it's assumed that any putatively responsible party has in fact committed an act for which he or she is possibly responsible in general free will terms, whether parsed by libertarianism or compatibilism. Additionally, I'd note here that my thought experiment could very well run in the opposite direction: we could start with deterministic rational beings verified via a D-detector who were thus assessed by compatibilist responsibility criteria, and then

later introduce comparably rational non-deterministic beings by mutation. There would still be no practical reason to assess the latter differently even if they came to be regarded as “angels” in miraculously choosing indiscernibly from D-beings (in fact given here the *base* view of compatibilism as “normal,” no one would *pragmatically* argue the *superiority* and *more-deserving* nature of indeterministically choosing beings). But in any case the obvious disadvantage is that that story couldn’t invoke the cool contemporary reference to zombies!

- 2 I owe Derk Pereboom much credit for this point, and in fact for raising issues in (2013) that gave birth to this whole zombie tale, though he would likely reject any sense that he ultimately deserves it.
- 3 Note that I allow that zombies are *only* rational. This simplifies the narrative and is within bounds of a possible world posit of the nature of the mutation.
- 4 Thus there is no empirical basis for claiming that libertarians are *disadvantaged* by potentially luck-bearing I-properties over zombies.
- 5 Of course genetic tests could disclose that they were mutants, and decisively so.
- 6 The late David Hodgson details how libertarian choice may be tied to choosing beliefs in (2012).
- 7 John Fischer (1994) has written on the nature of some such stalemates. Kristin Mickelson (this volume) has important reflections about *why* such stalemate dialogue occurs.
- 8 The *locus classicus* of this point is of course Susan Wolf (1980).
- 9 One distinct metaphysical facet of this complaint is about being *meta-unlucky*, which roughly is a form of luck that transcends any account of luck within a possible world: libertarians were meta-unlucky to inhabit a world where they would be treated differently as compared with their zombie companions. I intend to set out the significance of metaluck in a forthcoming paper (White). Still, most complaints about meta-luck would not be pragmatically based as is one focused on a one-world comparative case as the zombie example here.
- 10 This is why the zombie-world account assumes the Humean “black-box” problem of how we should understand that libertarian indeterminism is a claimed necessary condition of basic-desert responsibility given its *prima facie* pre-zombie correlation with reasoned choice, which then gives rise to the compatibility of the story with both event- and agent-causal forms of libertarianism since neither view has closed the deal (so far) on an account of that condition.
- 11 This is Gary Watson’s more eloquently expressed complaint about libertarianism in (2004). My zombies versus libertarians scenario just ups the morally pragmatic ante on the confrontation in one possible world.
- 12 What would turn the tide of argument back to the libertarian’s favor? Merely solve the “black-box” problem of how indeterminism specifically contributes to the ultimacy of basic-desert responsibility. Appeal to the I-detector can’t do it, even with the pre-zombie strict correlation of reason with I-properties, since that is undermined with the later appearance of equally reasonable zombies. Good (meta-) luck libertarians!
- 13 Saul Smilansky (2000) might well put his stamp of approval on this point in a general way with his illusionism. The zombie world arguably dispenses with the need for global illusion due to the I-detector, so that at least Humean-type libertarians might genuinely exist. Smilansky might then wish to argue that illusions need only be required for the equal treatment of zombies, though the resultant asymmetry of illusionism would at the least make for a morally very awkward world.
- 14 Another way to see the force of this parable? It is this: there is a strong epistemic parity between the world of its post-zombies situation and our world at present, particularly with respect to free will and responsibility theorizing. Not that the parallels are exact. For instance, much of the confidence about the truth of libertarianism at present in our actual world is not some form of correlation data from I-devices and human behavior, but rather the introspective and intuitive subjective evidence of feeling and sensing that we often possess an open-futured ability and at least dual-opportunity event of choosing in our lives – another kind of I-sourced evidence (if you will). Against that we have considerable evidence that we are not such indeterministically gifted creatures, whether we might be subject to mental disease or defect (as by Western insanity law, see Hartman’s chapter in this volume), or that more generally our decisions are not

ours to embrace, but the function of un-or-sub-conscious machinations of our brains. Just as the emergence of zombies played out against the strong evidence of I-devices in our thought-experiment parable, we have in our own real world clashing troves of evidence, both objective and subjective (though reversed in our world as against the zombie world), that put us in a collective place of not knowing exactly how to manage that evidence in terms of how we treat each other emotionally, morally, and legally – a place that I would call metaphysical epistemic depravity. The whole point of the zombie scenario is to portray our present epistemic dilemmas in a kind of reverse situation of such depravity. In that scenario we become objectively confident that most of us have libertarian free will, but then we confront those who subjectively confess that they are the same as those most of us, and objectively behave like us, but do not evince those I-device-demonstrated features. In our world, we (mostly) subjectively confess that we have libertarian free will, but the objective evidence that we do not has accumulated to an extent we cannot easily disregard. The strength of this parable resides in the fact that it shows in its way that we are and have been since time immemorial in a state of metaphysical epistemic depravity about how we exist as moral agents, given that we are moral agents at all. And as far as skepticism about our status as moral agents is concerned, the parable also advances that question as well – in the face of strong evidence that some people are libertarian agents and others behaviorally indistinguishable from them are not, reflective of the real epistemic metaphysical uncertainty of our own actual world about the general nature of agency, *should* we abandon all claims to responsibility as opposed to finding pragmatic grounds for assessing blanket descriptions of similarly comparable agents based on empirical evidence of rationally equivalent actors? While this cannot dismiss the skeptics' challenge to responsibility, it does place them in the position to respond to these pragmatically based claims that we can in fact undergird a system of justice with such evidence not just in our world, but an epistemically inversely posed zombie one.

- 15 As well as my coeditors Joseph Campbell and Kristin Mickelson, who significantly influenced my final presentation of this chapter with very helpful criticisms, I thank Derk Pereboom, Troy Cross, Manuel Vargas, and Neil Levy respectively for inspiration, clarification, support, and criticism. Levy notes in correspondence that this paper might be better presented as a straightforward argument for compatibilism, and I do not deny that as noted below. My point ultimately is that moral pragmatism as best enlisted as arbiter for the equal treatment of co-existing Libertarians and zombies could well be extended as an *axiological* basis for the preeminence of pragmatic compatibilist moral values over more absolutist incompatibilist ones not just in their world, but ours. This paper, preferring to focus on criticizing *moral* incompatibilism, is neutral about *metaphysical* incompatibilism.

Bibliography

- Fischer, J.M. (1994). *The Metaphysics of Free Will: An Essay on Control*. Oxford: Blackwell Publishers.
- Hodgson, D. (2012). *Rationality + Consciousness = Free Will*. New York: Oxford University Press.
- Pereboom, D. What are we fighting about? *Flickers of Freedom*. <http://agencyandresponsibility.typepad.com/flickers-of-freedom/2013/02/what-are-we-fighting-about.html> (accessed on 15th February 2013).
- Smilansky, S. (2000). *Free Will and Illusion*. New York: Oxford University Press.
- Watson, G. (2004). *Agency and Answerability*. New York: Oxford University Press, 184–196.
- White, V.A. (in preparation). *Luck and Meta-Luck*.
- Wolf, S. (1980). Asymmetrical freedom. *Journal of Philosophy* 77: 151–166.

Index

- ability, 3–4, 6–7, 10, 12–13, 61–64, 66, 69, 88–90, 95, 98–99, 104–105, 125–126, 129, 132–133, 157, 159, 162–163, 165, 169, 172, 175–176, 186, 196, 207, 210, 223, 225, 227–228, 235, 246, 250, 279, 316–317, 339–344, 346–349, 369, 373, 410, 419, 434–436, 441–442, 464, 472, 483
- conditional analysis of, 166, 212
- dual-, 98, 126
- specific, 125, 132
- strong/weak (or alternatives, can), 101, 104–105, 238
- to do otherwise, 3–4, 6–7, 10, 12–13, 61–64, 66, 69, 125–126, 165, 169, 172, 175–176, 186, 419, 441–442
- to refrain, 23–24, 61, 96–101, 104, 137, 164–165, 172, 175, 183, 411, 464, 468
- wide, 162, 168
- see also* capacity, control, power
- accountability, 343, 345–346, 359
- achievement, 115, 192, 197, 469–470
- action,
 - and* choice, 109, 112, 169, 192, 196, 223–224, 337, 378, 418, 426, 464
 - akrasia (akratic), 358–360
 - basic, 258
 - blameworthy, 359, 439–440
 - divine, 174
 - free, 6, 23–26, 28, 35–36, 40, 84–87, 89, 101–102, 118, 126, 135, 138, 146, 148, 151, 175, 188, 258–262, 264–265, 267, 270, 278, 280, 282–283, 287–288, 342, 380–383, 385–388, 457–458, 478–479, 481–482, 484–485, 488
 - future, 126, 211
 - human, 41–42, 44, 64, 86, 131, 138, 209–210, 269, 300, 380, 421, 464
 - immoral, 152, 342, 472
 - intentional, 266, 278–282, 284, 287–289, 346, 363, 382
 - past, 358, 371, 381–382
 - praiseworthy, 363
 - reasons for, 89, 113, 259, 261, 269, 288, 339, 360, 381
 - responsible, 482
 - self-forming, 63, 320
 - spontaneous, 280
 - theory, 1, 9, 11, 278
 - uncaused, 484
 - versus* omission, 100, 151, 357–359, 361, 379–380, 384, 387, 406
 - versus* torn decisions, 320
 - zombie, 266
- addiction, 340, 408, 426
- agency/agent,
 - autonomous, 40, 53
 - designed, 147
 - divine, 108, 113
 - experience of, 265, 282, 419–430
 - free, 63, 71, 90, 92, 146–147, 176, 182, 192, 213, 281, 386–387, 417–430, 460
 - human, 40, 42–44, 53, 108–111, 114, 118–119, 150, 152, 211, 230, 308, 310–311
 - intentional, 180, 437
 - moral, 10, 337, 349
 - neuroscience of, 279, 283, 288
 - phenomenology of, 92–93, 258, 281, 460, 488
 - responsible, 180, 190, 192, 195, 199, 205–206, 211, 296, 337–338, 349, 417, 488
 - sense of, 200
 - theories of, 241, 262
 - theory, 92
 - voluntary, 114
- agent-causation, *see* causation
- akrasia (akratic), 358–360
 - belief, 358–360
- alterability, 96, 106, 171
 - see also* unalterability

- alternative possibilities (or alternatives),
 23, 95, 99, 101, 125, 210, 386–387,
 441–442, 464
 conditional analysis, 135, 212
 see also ability (to do otherwise), principle of
 alternative possibilities (PAP)
 alternatives, 184, 188–189, 196–197
 viability of, 184, 188–189, 196–197
 anger, 212, 340, 345, 349, 351, 370, 372–376,
 393, 398, 402, 404, 472
 Anselm, 108–110, 119, 170
 see also reactive attitudes
 answerability, 343–346
 Aquinas, 108, 110, 169–170, 294
 Aristotle, 109–112, 241, 338, 349, 355
 attitudes,
 affective, 398–399
 -based reasons, 399–400
 constitutive, 360, 402
 emotions/emotional, 208, 369–377
 implicit (vs. explicit), 9, 241–246, 248–252
 interpersonal, 370
 pro-, 401
 propositional, 204, 207, 242, 264
 reactive, 193–196, 198, 294, 296, 343,
 375–376, 398, 470–472
 responsibility-based, 197
 responsibility-characteristic, 213, 216
 second-order, 316
 attributability, 248, 251, 343–344, 346, 359
 Augustine, 109–110, 176
 autonomy, 41, 115–119, 191
 differentiated, 115–119
 avoidability, 97, 99, 104, 144, 150–151, 152

 basic argument, 8, 13, 65–66, 70–71, 186,
 381–383, 387, 477
 basic desert, *see* desert
 bivalence, 6, 14, 24–26, 33–35, 160
 blameworthiness, 10, 95, 98–101, 104, 152,
 194, 213, 244, 296, 309–310, 337–338,
 346, 348, 355, 357–366, 372, 380–381,
 284–287, 396–397, 401–402, 410,
 439–440, 465, 492
 see also praiseworthiness
 body, 165, 225, 229, 258, 268, 278–280,
 305, 479
 Boethius (or Boethianism), 30–32, 36, 171
 bypassing, 300–311, 469

 can, *see* ability
 capacity,
 causal, 85
 cognitive, 115, 411
 diminished-, 412
 narrative-, 347–348
causa sui, 66, 186, 205
 causal,
 account, 41–44, 85–87, 370, 382–383, 418
 agent-, 5, 85–93, 319, 323, 325, 370, 445,
 467–469, 478–479, 484–487, 491
 and nomological, 2–3, 14, 65, 126, 130, 132,
 137, 236
 capacity, 85
 -control, 53, 130, 324–325
 -dependence views, 257–261, 265, 267–270
 determinism, 8, 40–44, 53–54, 84, 147, 246,
 376, 379, 464–465, 469, 474
 efficacy, 270, 284, 453
 event-, 5, 84–85, 88, 90–93, 126, 205,
 247–248, 318, 321, 324, 369–370, 376,
 467–469, 484–488, 491
 factors, 3–4, 7, 13, 63, 209, 469
 history, 130, 148, 162, 164–165, 443
 indeterminism, 6, 91, 370–371, 380
 luck, 3–5, 11, 13, 14, 65–66, 68, 70–72, 78,
 135, 379–380, 384, 386–387
 models, 306–307, 427–430, 438, 445, 447
 non-, 50, 52, 65, 68, 84–85, 181, 319
 power, 85, 88–92, 223–227, 230, 235–236,
 282–283, 418, 468, 478–479,
 484, 486–487
 relation, 92, 126, 164–165, 264, 286,
 269–270, 438, 442–443, 447
 causation (and free will), 2, 8, 14, 63, 65, 67,
 85, 88–91, 136, 223–224, 228, 319, 324,
 370, 434–448, 468, 477, 484, 491
 agent-, 85–93, 180, 319, 323–325, 370,
 382–383, 418, 445, 447–448, 477–479,
 484–487, 491
 backward, 171
 event-, 85, 87, 89–90, 319
 Humean, 14, 42–43, 68, 131–132, 170,
 235–236, 491
 interventionist, 434–435
 mental, 260, 438
 metaphysics of, 434–436
 probabilistic, 8, 378
 substance-, 85, 88–90, 369, 467–469
 see also libertarian accounts of free will
 children, 10, 113, 174, 193, 244, 337–349,
 372, 375, 470
 see also moral responsibility (children)
 choice, *see* free choice
 classical analytic paradigm (CAP), 5, 64, 67, 69,
 72

- classical analytic paradigm (CAP) (*cont'd*)
 - and mysterianism, 7, 140
 - and the classical bridge inference, 62, 63, 67, 68, 76, 140
 - and the "classical" consequence argument, 7, 75–76, 125–140
 - defined, 3, 12, 58–62
 - degeneration of, 7, 62–63
- classical compatibilism, *see* compatibilism (classical)
- classical incompatibilism, *see* incompatibilism (classical)
- cognition, 241, 243, 246, 248, 294, 339, 428
 - type-one *versus* type-two, 243, 248
- collaborativism, 211
- compatibilism,
 - actual-sequence, 441–444
 - agnosticism about, 289, 485
 - and moral responsibility, 4, 13, 15–16, 55, 68, 76, 146–149, 180–181, 184–188, 193, 195, 197–199, 205–206, 210–213, 223, 246–247, 251–252, 281, 295–302, 308–311, 338, 364, 376, 378–379, 381, 383–385, 387–388, 394, 417–419, 421, 424, 435–436, 439–440, 445, 465–467, 481–484, 491–493
 - anti-classical, 16–17
 - classical (ability-to-do-otherwise/leeway), 3–4, 12, 14, 16–17, 62, 140
 - defined/disambiguated, 3, 12, 14, 16–17, 39, 58–83, 140, 216, 237–238, 247, 281, 465, 474, 481, 488
 - history-sensitive, 381, 383
 - natural, 294–313
 - neo-classical, 4–5, 16–17
 - post-classical, 16–17
 - revisionist, 198, 212
 - semi-, 15–16, 76
 - theological, 180
 - see also* compossibilism, incompatibilism, classical analytic paradigm (CAP)
- compatibility relation, 67–72
 - see also* compossibility relation, consistency relation, incompatibility relation
- compatibility problem/question, *see* problem of (free will) and determinism
- compossibilism, 3–5, 7, 9, 13, 16–17, 60, 62, 66–67, 72–73, 75–77, 474
 - classical, 66
 - source, 76
 - see also* compatibilism, impossibilism
- compossibility relation (metaphysical), 67–72
 - see also* compatibility relation, impossibility relation
- compulsion (or compulsive), 181, 186, 209, 260–262, 320, 411
- conditional analysis, *see* ability
- conditional analysis argument, 135, 140
- consciousness, 51, 249, 252, 256–270
 - see also* mind
- consequence argument, 7, 13, 28, 35, 42, 53–54, 125–140, 228
 - classical, 7, 13
 - inevitability version, 129–135
 - no-past objection, 130
 - unalterability version, 133–139
- consistency relation (logical), 15, 71, 79
- constraint, 7, 169–172, 176, 188, 191–192, 196–198, 205, 227–228, 281, 287, 340, 342, 361, 393
- contextualism (or context-sensitivity), 158–159, 163, 166, 187–188, 210, 248, 349, 426–427
- control, 3–8, 10–11, 13, 15–16, 40–41, 43–44, 46, 49, 53–54, 59, 63–67, 70–71, 84–85, 87–92, 97–104, 109–112, 118–119, 130–132, 134–135, 137, 141, 144–149, 169, 172, 174–175, 179, 182, 185–186, 188, 193–195, 199–200, 224, 241, 243, 248–25, 257, 259–267, 269–270, 278, 279, 281, 283, 286–287, 294–295, 299, 301, 305, 314, 316, 322–326, 338–341, 343, 346, 355–357, 364, 369, 371, 376, 378–385, 388, 400–401, 406, 410–411, 417–419, 434, 438, 441–443, 464–470, 473, 483, 485–487
- active, 84, 485
- agent, 92, 263, 314, 371, 376, 467
- and free will (or freedom), 13, 40–41, 43–44, 53–54, 59, 63–67, 70–71, 84–85, 88–92, 101–103, 109–112, 118–119, 130–131, 134–135, 144–149, 169–170, 172, 175, 185–186, 188–189, 193–195, 199–200, 241, 248–251, 257, 261–262, 264–265, 269–270, 278, 279, 281, 294–295, 305, 314–315, 338–339, 341, 343, 355, 369, 378–392, 417–419, 434, 441–442, 464–465, 467–469, 485–488
- and God (or gods), 5, 8, 10, 13, 15, 40–41, 70, 109–112, 118–119, 146, 169–170, 172–175, 179, 182, 186, 326
- and moral responsibility, 10, 13, 188, 241, 248–251, 257, 259–261, 263, 279, 281,

- 294–295, 338–339, 379, 383, 401,
417–418, 434, 438, 464–465
- beyond our, 3–6, 8, 59, 63–67, 70–71,
84–85, 87, 100, 130–131, 134–135,
193, 195, 378–380, 384–385, 465–467,
469–470, 473
- causal, 53, 84, 130, 324–325
- condition, 10, 13, 248–249, 257, 259–261,
263
- conscious, 260, 269, 286
- direct/indirect, 90, 174, 260–261,
485–487
- enhanced, 85, 87–88, 90, 92, 314–316
- executive, 341, 350
- guidance, 272, 485
- over the laws, 5, 66
- over the past, 5, 7, 66, 84, 103, 112
- over what comes to mind, 103
- plural, 314
- providential, 109, 112, 169
- regulative, 15, 485
- self-, 146, 174–175, 186, 340, 343, 349,
410–411, 483
- ultimate, 10, 186
- see also* problem of enhanced control, free will,
moral responsibility (basic desert)
- correlation problem, *see* problem of determinism
- counterfactual dependence, 231, 257–258,
261–263, 269
- counterfactuals, 162–163, 165, 181, 223–225,
228–236, 248, 286, 429
- criminal responsibility, 154, 406–412
see also culpable (culpability)
- culpable (culpability), 154, 406–412, 338,
355–359, 365–366, 372, 412
- ignorance (*versus* nonculpable ignorance),
338, 355–359, 365–366
- decisions, 40, 53, 84–94, 102–103, 135–138,
145, 165, 173–175, 181, 194, 209, 224,
228, 230, 232, 235, 247, 278–293, 297,
300–301, 305, 314–334, 344, 369, 429,
436–440, 443, 466–472, 477–481, 487,
488, 493
- and* God, 173–175, 181
- conscious, 9, 279–293, 322, 324, 327, 329,
480–481
- decision-making, 89, 228, 247, 279,
288, 315–318, 418
- free decision, 91, 102, 138, 278, 282, 315,
319, 328, 468, 471
- moral, 344
- proximal (or intentions), 286
- rational, 493
- torn, 319–330
- deliberation, 52, 89, 109, 126, 145–146, 191,
206, 209–210, 263, 270, 272, 274, 288,
345, 380, 417–418, 429, 467, 469–470,
483
- rational, 483
- denialism, 1–2, 5–6, 10, 11–12, 13–14, 15,
73–74, 184–203
- absolute, 187, 198, 200
- partial, 187
- see also* free will skepticism
- Dennett, Daniel, 15, 65, 68, 74, 78, 93, 216,
319, 419
- desert, 10, 13, 64, 76, 184–203, 206,
215–216, 244, 309–310, 325, 371–374,
376, 388, 393–405, 418, 465–468, 470,
487, 491, 493
- base, 395–399, 418
- basic, 10, 13, 64, 76, 186, 190, 194, 200,
244, 325, 374, 376, 388, 393–405,
465–468, 470, 487, 493
- just, 10, 186, 189
- moral, 206
- sense (of moral responsibility), 190,
371–374, 465–466
- see also* moral responsibility
- desire, 85, 103, 109, 137, 145, 159, 172–173,
175, 180, 186, 244, 246, 247, 262, 263,
279, 298, 300, 304, 315–316, 340,
342–343, 420, 452, 466, 473, 481–483
- and* God, 172–173, 175, 180
- children's, 340, 342–343
- first-order, 316, 440–441
- higher order, 340
- irresistible, 340
- second-order, 440–441
- determinism
- causal, 8, 40–44, 53–54, 84, 147, 246, 376,
379, 464–465, 469, 474
- divine hard, 181
- E-causal, 41–44, 52, 54–55
- hard/soft, 2–3, 13, 73, 181, 184–185, 380,
465
- Humean, 14, 77
- Laplacian, 55
- logical, 40, 300
- M-causal, 41–44, 50, 55
- ontological, 111–112
- neurological, 300, 303
- pre-, 6, 14

- determinism (*cont'd*)
 - providential, 6, 473
 - psychological, 300, 310
 - R-causal, 44, 49–51, 53–54
 - related factors, 3–5, 13–14, 59–65, 67–68, 71, 73
 - socio-economic, 6, 14
 - sociological, 40
 - theological, 40, 473
 - traditional, 2, 12–15
 - see also* fatalism, indeterminism (traditional), inevitability (inevitabilism)
 - problem of determinism
- deterministic laws, 44, 70, 223, 226, 229, 281, 468
- dialectical stalemate, 5, 492
- direct argument, 144, 149–153
- disabilities, 10, 343, 347, 410–412
- disappearing agent argument, 468
- discretionism, *see* revisionism
- dispositions, 90–91, 194, 208, 210, 242, 245, 344, 360, 364, 373, 379, 381–384, 387, 434, 470–471
 - and* constitutive luck, 383, 387
 - emotional, 344
 - lucky, 383, 387
 - mental, 381–382, 384
 - to blame, 245
 - see also* luck (constitutive)
- dualism, 52, 229, 478–479
 - of mental and physical properties, 281
- Einstein, Albert, 11
- eliminativism, 204, 212–216, 393, 401, 451–463
 - see also* denialism, skepticism (about free will), revisionism
- emotions (moral), *see* attitudes
- empiricism, 11
- epistemic condition on moral responsibility, *see* moral responsibility
- equability, 139–140
- eternal, 30–32, 36, 108, 112–113, 116, 169–170, 181, 297, 301, 308
 - recurrence, 297, 301, 308
- eternalism, 6, 14, 29–32, 36
- events, 23–24, 28–31, 36, 84–94, 96, 99, 125–128, 131–134, 137–139, 147, 149, 158, 160, 163–165, 205, 209, 225–226, 229–230, 232, 246–248, 257, 259, 264, 266, 282, 285–287, 300, 302, 304, 314, 318–324, 326, 347, 349, 369–371, 376, 380, 388, 428, 435–438, 440, 442–443, 464–465, 467–469, 479, 483–489, 491
 - concrete, 23–24, 28–31, 36
 - mental, 51, 53, 323
 - see also* event-causal libertarianism
- evil, 34, 110–111, 113, 118
 - see also* problem of evil
- experimental philosophy/philosophers
 - (X-Phi), 9, 11, 12, 17, 199, 205, 207, 210, 245, 267, 295, 298–299, 478–479, 487
 - see also* Libet experiments, compatibilism (natural), neuroscience, revisionism, X-Phi
- explanation,
 - abductive (inference to the best explanation) reasoning, 17, 75, 77, 78, 153, 467
- action, 44, 181, 209–210, 250, 328
- causal, 54, 224, 338, 439
- event-based, 54
- explanatory gap, 63
- explanatory-relevance relation, *see* relevance relation
- explanation problem, *see* problem of determinism
- facts, 2, 5, 28–29, 31, 46, 61, 63, 70, 101–102, 104–105, 112, 132, 141, 157–170, 175, 182, 185, 229, 231, 234, 242, 244, 248, 316–317, 325, 328, 345, 348, 356, 361–362, 364, 397–398, 400, 402, 442–443, 485
 - about actual causal histories, 442
 - about the distant past, 102
 - about the early state of the universe, 70
 - about the past, 2, 61, 63, 105, 229, 234
 - about the past and the laws, 101, 104–105, 141, 229
 - all the, 170, 175
 - contingent, 182
 - empirical, 316
 - external to the causal history, 443
 - hard, 28–29, 31
 - metaphysical, 402
 - non-normative (non-moral), 356, 361–362
 - of logic and mathematics, 169
 - physical, 170
 - relevant, 158–159, 317, 364
 - settled, 132
 - soft, 28–29, 31
 - see also* propositions
- fatalism, 23–36, 160–162, 165, 226
 - logical, 24, 26–28, 30, 32, 34–36, 160–162, 165
 - theological, 24–28, 30–32, 34–36

- Fischer, John Martin, 64, 68, 100, 128–129, 147, 172, 186, 241, 248–249, 279, 339, 342, 357, 381–383, 385, 387, 419, 440, 466, 481, 485
 and Mark Ravizza, 64, 172, 186, 241, 248–249, 279, 339, 342, 381–383, 385, 419, 440, 481
- fixity, 8, 26–31, 35–36, 40, 164, 171
 of the eternal, 31–32, 36
 of the past, 26–31, 35–36, 164, 171
 of the present, 26–31, 35–36
- foreknowledge, 6, 14, 25, 70, 79, 111–112
see also omniscience, theological fatalism
- foreseeability, 356–357
 of ignorance, 356–357
- forgiveness, 10, 369–377
see also reactive attitudes
- Four-case Argument (Plum), 7–8, 14, 17, 78–79, 102, 145–149, 394, 466–467
see also manipulation (multiple-case) argument
- Frankfurt examples/cases (Black/Jones), 4, 58, 95, 98–101, 104, 294–295, 419, 441–445, 447
- Frankfurt, Harry, 4, 58, 95, 98–101, 104, 172, 186, 294–295, 316, 340–341, 419, 440, 441–445, 447, 481–482
- Freakish Demon Argument, 6
see also manipulation (multiple-case) argument
- free action, 6, 23–26, 28, 35–36, 40, 84–87, 101–102, 126, 135, 138, 146, 148, 151, 175, 188, 258–262, 264–265, 267, 270, 279, 280–283, 287–288, 342, 380–383, 385–387, 457–458, 479, 481–482, 485, 488
- free choice, 86–87, 90, 110–113, 138, 169, 181–182, 195, 230, 284, 382, 386–388, 410, 426, 458, 468
- freedom,
 as ability and opportunity (-ies), 105, 159, 250, 348, 494
 concept of/conceptions of, 63, 65, 241, 257–260, 316–317, 338, 379–381, 426, 442, 444, 448
 divine/foreknowledge, 25, 70, 112, 169, 175–177, 181–182
 from constraint, 172, 190–191, 397
 implicit attitudes, 244, 246–250
 luck, 379–384, 386–387
 metaphysical, 60, 62, 67, 92, 102, 445
 moral (responsibility), 8–9, 92, 96, 136, 147–148, 180–181, 186–187, 192, 195–196, 205, 213, 243, 339, 355, 379, 434
 of action, 173, 226, 241, 484
 of the mind, 89, 108–109, 171–172, 173, 244, 268–270, 329, 425
 perfect (God's), *see* perichoretic freedom
 religious, 108, 111, 471
 time-travel, 157, 164
 to do otherwise, 61, 95–98, 103, 110, 113, 165, 176, 186, 314–315, 321, 324–325, 386–387
- free will,
 actual causal sequence view of, 62, 441–445, 447
 agent-causal account, 5, 55, 75–76, 85–93, 319, 323, 325, 331, 370, 382–383, 418, 430, 448, 467–469, 478–479, 484–487, 491, 494
 agnosticism about, 289, 485
and the control condition on (basic desert)
 moral responsibility, 10, 185, 200, 244, 325, 374, 376, 388, 464–465, 486–487
 classical (ability-to-do-otherwise/leeway)
 account, 3–4, 7, 12–13, 15, 23, 36, 61–62, 65, 74, 77, 125–126, 128, 134, 139–140, 176–177, 186, 338, 464
 defense (and the problem of evil), 110–111, 113, 133, 172, 180–182, 472
 event-causal account, 5, 55, 75, 84–85, 88, 90–93, 126, 205, 247–248, 314, 319, 322, 324, 331, 369–370, 376, 448, 467–469, 484–488, 491
 fixing the referent of, 1–2, 10, 13, 16, 63, 204–219, 451–463
 free-will thesis, 6, 12–14, 58, 484
 neuroscience of, 278, 280, 282, 288
 of God, 175
 revisionism, 8, 11, 198, 204–215, 376, 394, 400, 456, 458
 robust, 185–186, 188, 429, 430, 441, 468
 skepticism, 15, 125–127, 134, 138–140, 184, 189, 191, 281, 369–376, 430, 465, 469, 471, 487, 491–492
 source account, 4–5, 10, 12, 125–126
 sourcehood, 4, 62, 66–67, 103, 125, 172, 210, 338, 435, 437, 439
 zombies, 492–493
see also problem of free will
- free-willism, 1–2, 12–14, 125–127, 130, 136
- freedom, *see* control, free will, L-freedom
- funishment, 190
see also punishment

- Ginet, Carl, 41, 127, 128, 258, 321
- God, 25–27, 30–32, 40, 70, 108–121, 169–183, 175, 182, 225
 intentions/will of, 173–175, 180–181
 see also foreknowledge, omniscience, theological fatalism
- grandfather paradox, 157–158
 see also time (travel)
- gravity, 46
- grounding principles, 131
- guidance control, 485
 see also regulative control
- guilt,
 actus reus (guilty act), 406–407
 and the presumption of innocence, 201
 criminal law *versus* tort law, 407–409
 criminal *versus* moral, 409
 feelings of, 195, 482
 mens rea (guilty mind), 366, 406–407
 see also reactive attitudes
- hard determinism,
 defined, 2, 73, 465
 divine hard determinism, 181
 versus denialism, 13, 73, 184, 185
 versus hard incompatibilism, 380, 403
 versus libertarianism, 2, 465
 versus soft determinism, 73, 185, 465
- hard incompatibilism, *see* incompatibilism
- hard facts,
 versus soft facts, 28–29
 see also Okhamism
- hell, 10, 113, 180
- hope, 472–474
- Humeanism (causation/laws of nature), 14, 42–43, 45, 68, 77, 131–132, 169–170, 235–236, 491
- identity, *see* personal identity
- ignorance,
 and the epistemic condition on moral responsibility, 10, 184–203, 355
 culpable, 358
 exculpating *versus* inculpating, 355
 moral, 361–362
 see also knowledge, volitionism
- illusionism, 11–12, 94, 199, 494
 see also revisionism
- immutability (or immutable), 26–27, 32
 knowledge, 27
 principle, 26, 32
- implicit bias, 243, 246–248
- impossibilism,
 anthropocentric restrictions, 12, 75, 125, 403, 430
 arguments for, 7, 62, 139
 defined, 13, 16, 62
 explanations for, 78
 versus compatibilism and incompatibilism, 16–17
 versus hard incompatibilism, 4–5, 62, 75–76, 78
- inability, 426
- incarceration, 190–193
 see also punishment
- incompatibilism, 58–83, 297, 369, 483
 ambiguity of term, 15, 16, 54, 58, 66–67, 78, 153, 388, 474, 488
 anti-classical, 16–17
 as solution to the explanation problem, 68, 72
 broad, 144, 146, 153
 classical (ability-to-do-otherwise/leeway), 3–4, 5, 7–8, 13, 16, 62, 66, 108, 224, 228, 464
 classical bridge inference (from impossibilism), 62
 defined/disambiguated, 13, 39, 54, 58–83, 84, 125, 153, 184, 237, 246–247, 281–282, 289, 388, 418, 474, 481, 488
 hard, 13, 75, 184, 380, 401, 456, 458, 492
 history of term, 13, 15, 38, 73
 natural, 295
 neo-classical, 4–5, 16–17
 post-classical, 16–17
 source, 4–5, 7, 111, 113, 118
 virtue-, 108
 see also compatibilism, impossibilism, impossibilism, problem of determinism
- incompatibility relation, 3, 6, 28, 53, 58, 67–72, 80, 98, 126, 140, 149, 150, 151, 152, 153, 164, 289, 388, 484
 see also inconsistency relation, impossibility relation
- impossibilism, 7, 8, 16–17, 58–83
 classical, 61, 66, 69, 76
 classical bridge inference (to incompatibilism), 62
 defined, 13–14, 60
 history of term, 73
 rival explanations for, 63–64, 66–67, 474
 source, 62, 72
 see also compossibilism, incompatibilism, impossibilism
- impossibility relation (metaphysical), 60, 69–72, 79, 383

- see also* inconsistency relation, incompatibility relation
- inconsistency relation (logical), 15, 71, 79
- indeterminism, 84, 97, 209, 306, 418, 426
and luck, 135–136, 318, 380
and physics, 8, 42–44, 318
as umbrella term, 25, 185, 319–320, 464, 465
 defined, 2, 14, 50, 59, 319–320, 370
 R-indeterminism, 50–52
 traditional, 2, 14, 77
see also determinism, Humeanism, laws
- inevitability, 2, 14, 59, 61, 129–134, 148, 196, 347, 463, 469–470, 473
- inevitabilism, 74
- version of the consequence argument, 129–135
see also determinism
- inference to the best explanation, *see* *explanation*
- insane/insanity, 145, 376, 492
 insanity defense, 406, 408–411
see also sanity
- intent/intention/intentional, 84–86, 90, 102, 105, 112, 117, 148, 150, 173, 175, 180–181, 192, 244, 258, 363, 365, 379–382, 393, 409, 437, 438, 435, 466, 471, 480, 481
and consciousness, 258–267, 269, 420
and children, 341, 346
and forgiveness, 370–371
and mens rea, 406–407
and revisionism, 208
see also God, Libet experiments
- intuition
 causal (morally), 441–442, 444–448
 children's, 341–342
 folk, 9, 12, 60, 244–246, 251–252, 294–299, 308–309, 318, 338, 383, 457, 460
 incompatibilist, 224
 impossibilist, 67
 pump, 74
see also manipulation (multiple-case) arguments, revisionism, X-Phi
- involuntariness, *see* voluntary
- irrational/irrationality, 145, 245, 246, 268, 374, 376, 466
- judgment,
 by children, 399, 344–345, 348
 causal, 230, 407, 439
 divine, 111, 115
- in forgiveness, 371–372
- introspective experience, 420, 425
- manipulation cases, 144, 147, 444, 447, 482–483
- moral, 204, 207, 210, 216, 298–303, 339, 374, 387, 400–401, 438, 446
- reference-related, 459–460
- justification, 8, 192, 215, 325, 376, 385, 395, 398–400, 407, 467, 493
- causal, 324
- fairness, 400, 402
- for blame, 174, 398, 400 (*see also* blame)
- for punishment/retributive treatment, 191, 193, 325, 493
- for skepticism, 467
- pragmatic, 376, 493
- Kane, Robert, 6, 9, 15, 55, 59, 63, 64, 75, 102–103, 126, 135, 138, 172, 185, 247–248, 289, 314, 320, 321–322, 330, 331, 369, 370, 382, 387, 430, 431, 448, 466, 467, 481, 485–486
- Kant, Immanuel, 118
- knowledge
 counterfactual, 225, 229, 231
 criminal law, 406–407, 409–410
 epistemic condition on moral responsibility, 338, 342, 347, 356–357, 362–363, 382–383
 immutable, 27, 34
 self-knowledge, 250, 342, 349
see also omniscience, ignorance, *mens rea*
- laws
 causal, 65, 70
 of nature, 6, 7, 41–43, 52, 59, 64, 65, 70, 88, 91, 96, 105, 126–135, 137–138, 141, 146, 149, 150, 151, 154, 163, 169, 224, 226–227, 247, 257, 259, 281, 296, 320, 369, 370, 384, 435, 464, 467, 469, 470, 483, 485
 physical, 43, 48, 88, 89, 468
see also Humeanism, determinism (Humean)
- leeway (in actual sequence), 3, 7, 13, 15, 61–62, 65, 74, 77
- God's, 176–177
see also free will (classical accounts), indeterminism, inevitability
- Lehrer, Keith, 3, 15, 73, 78, 131, 141
- liability, 406–408
- libertarianism (libertarians), 9, 11, 314–334
 defined, 2, 126, 484

- libertarian accounts of free will, 314–334
 - agent-causal accounts, 5, 76, 85–93, 319, 323, 325, 331, 370, 382, 383, 418, 467–469, 478–479, 484, 486–487
 - classical (ability-to-do-otherwise, leeway) accounts, 3, 6, 10, 12, 61–62, 126, 165, 172, 441
 - event-causal accounts, 5, 84–85, 90–93, 126, 205, 247–248, 331, 369–370, 376, 467–469, 484–488, 491
 - non-causal accounts, 84, 319, 331
 - see also* libertarianism, classical analytic paradigm (CAP), free will
- Libet experiments, 9, 266–267, 273, 279–287, 477–478
- vetoing, 273, 282, 289
- love, 113, 117, 119, 193–195, 197, 208, 214, 370, 470–471
 - see also* reactive attitudes
- luck, 12, 55, 66, 77, 91–93, 97, 100, 135, 141, 182, 193, 227, 247–248, 269, 274, 318–321, 323, 325–326, 356, 361, 364, 369, 388–392, 487, 494
 - and* torn decisions, 319–320
 - causal, 3–5, 11, 13–15, 65–66, 68, 70–72, 76, 78–79, 135, 379, 384, 386–387, 389, 443
 - causal/circumstantial hybrid, 66
 - chancy, 318
 - circumstantial, 65, 379–380, 383–387
 - consequential/resultant, 364–365, 379–380, 385–389
 - constitutive, 5, 8, 11, 13–15, 17, 65–67, 70, 76–78, 364, 369, 379–387, 389
 - cross-world, 91
 - factors beyond one's control, 3, 6, 76, 380
 - indeterministic, 136
 - Luck Pincer, 8, 381–383, 387
 - metaluck (*see also* cross-world luck), 494
 - non-causal, 17, 65
 - present (problem of), 90–93, 106
 - principle, 100–101
 - Type-2, 443–444
 - see also* moral luck
- manipulation, 102–103, 180–181, 248, 304–305, 315, 375, 381, 441, 443–444, 447, 481–483
- manipulation argument (a.k.a. multiple-case argument), 6–8, 40, 74, 144–156, 130, 294, 295, 394, 434–439, 466, 471, 484, 488
 - see also* Four-Case Argument, Freakish Demon Argument, Master Manipulation Argument, Zygote Argument
- Master Manipulation Argument, 8, 17, 74
 - see also* manipulation (multiple-case) argument
- materialism, 318–319, 323, 326, 452–453
 - see also* physicalism
- McKenna, Michael, 5, 68, 78, 141, 146–147, 151–152, 154, 196, 215, 217, 339, 380, 389, 394–395, 397–398, 403, 419, 435, 436
- meaning in life, 185, 193–197, 198, 200, 370, 464–476
- mechanism, *see* reason-responsiveness
- Mele, Alfred, 7–8, 16–17, 42–43, 53, 74, 76, 77, 78–79, 86, 90–91, 103, 146–147, 154, 180, 247, 267, 272, 273, 274, 286, 289, 290, 294, 319, 331, 340, 346, 379, 380, 382, 383, 389, 418, 419, 435, 466, 477–490
- mens rea* (*mentes reae*), 366, 406–407, 412
- mind, 9, 40, 49, 51–53, 86, 89, 228, 242, 248, 246, 294, 300, 303, 305, 322, 326, 328, 341, 381, 386, 478–479, 480
 - of God, 30, 176
 - see also* causation (agent-causal, mental), consciousness, *Mind* Argument, phenomenology
- Mind* argument, 6–7, 125–126, 130, 133–140, 318
- Molinism, 111–113, 181
- monism, 187, 198, 338
- moral emotions, *see* anger, blame, hope, love, reactive attitudes, resentment
- moral luck, 10, 64–65, 75, 76, 77, 78, 378–392
 - see also* luck
- moral responsibility, 4, 10–11, 13, 63–64, 68, 84, 91, 95–96, 98, 101–103, 144–153, 154, 180–182, 184–199, 205–206, 210–211, 213–216, 223, 241, 243–246, 252, 258, 279, 288, 294–303, 305–307, 308–311, 316–318, 325, 337–349, 355–357, 359–366, 372–374, 379–388, 393–394, 396–397, 401, 409–410, 418–419, 424, 434, 436–438, 440–441, 445–447, 457, 464–468, 470–471, 478, 481, 484, 487–488, 491–493
- degrees of (mitigation of), 1, 100, 101, 148, 251, 338, 346, 352, 363–365, 380, 384–388, 389, 439–440, 445, 447, 448

- epistemic condition on, 10–11, 206, 241, 244, 338, 345, 348, 355–368, 448
- free will as control condition on, 10, 13, 63, 76, 84, 248–249, 257, 259–261, 263
- heaven/hell (heaven-and-hell), 10, 113, 180
- of children, 10, 113, 244, 337–354, 375
- of the mentally ill, 343, 347, 410–412
- without free will, 15, 76, 194
- see also* blame, praise, blameworthiness, desert (basic), forgiveness, praiseworthiness, moral luck, responsibility
- mysterianism, 75, 127, 140, 492
- defined, 7
- Nagel, Thomas, 65, 379–380
- naturalism,
- naturalistic philosophy, 318–319, 326
- nomological, 126, 132, 134, 137
- negligence, 190, 358
- criminal, 406–407
- neo-classical compatibilism, *see* compatibilism
- neo-classical incompatibilism, *see* incompatibilism
- no-choice operator (N-operator), 128–131, 138–139
- no-difference claim, 145–147, 153
- no-past objection, 130
- non-responsibility, 382, 466, 467
- non-responsibility operator (NR-operator), 149, 151–152, 154
- neuroscience, 9, 11, 54, 278–393
- normativity, 95, 198, 206, 210, 212–213, 215–216, 249, 252, 345, 356, 394–395, 399–400, 402, 407, 410, 438–439, 445, 447
- nudge/nudging, 389
- obligation,
- moral, 95–107, 362–363, 365, 372
- epistemic, 358–360, 366
- Ockham/Ockhamism, 28–29, 31, 36, 109–110, 112–113
- see also* foreknowledge, fixity of the past, Molinism
- omission,
- moral responsibility for, 100, 151, 357–359, 361, 379, 380, 384, 406, 442–445
- “Frankfurt Case-Omission”, 444
- omnipotence, 35, 110, 111, 113, 118, 169–170, 175–177
- see also* God, voluntarism, omniscience
- omniscience, 6, 25–7, 34, 36, 175
- essential, 25–7, 34, 36
- see also* foreknowledge
- openness (epistemic), 84, 429–430
- open theism, 32–35, 112
- opportunity (and freedom), 105, 158–159, 348, 364, 379, 384–385, 494
- for reflection, 243, 250
- see also* ability
- origination, 86, 102–103, 111, 153, 172
- ought, 95–96, 98, 104–105, 346, 348–349
- implies can, 98–99
- see also* principle of alternative possibilities (PAP)
- overdetermination, 151–152, 232–233, 446
- ownership, *see* reasons-responsiveness
- paradigm case argument, 135
- past, 2, 29, 43, 50, 61, 91, 96, 97, 101, 104, 105, 128, 130–133, 141, 150, 158, 164–165, 170–171, 181, 194–195, 224, 229–236, 247, 297, 467, 485
- alteration of, 160–163
- distant, 84, 102–103, 149, 369, 371
- hard and soft, 28
- local, 49
- mutability of, 158
- Principle of Fixity, 26–28, 30, 31, 35–36, 164
- remote/distant past, 5, 7, 42, 64, 66, 84, 102, 103, 126, 128–129, 130–139, 137–139, 149, 230, 369, 371, 464, 469–470
- Pereboom, Derk, 7–8, 62, 76, 78–79, 84, 88–89, 93, 102, 118, 138, 145, 147, 189–191, 194, 198, 200, 216, 244, 323–325, 349, 370–372, 374–376, 387, 388, 394–396, 403, 418, 430, 456, 464–467, 477, 485, 486, 487
- perichoretic* “perfect freedom”, 116, 118
- personal identity, 64, 77, 204, 338, 347
- personal relationships, 369–370, 376, 471–472
- phenomenology, 11, 92, 419–427, 460, 488
- physicalism, 118, 480
- see also* materialism
- physics, 8, 11, 42, 48, 88, 90, 118, 141, 208, 212, 227, 241, 260, 318, 468
- Newtonian, 8
- quantum, 6, 8–9, 42, 44, 49, 51, 53–54, 133, 141, 247, 260, 323, 326
- pluralism, 187, 199, 338, 447–448
- possible worlds, 128, 130, 159, 177, 207, 227, 229–231, 236, 488, 492
- power, 85, 88–92, 97, 109, 116, 144, 150, 169–171, 175, 177, 186, 223–227, 236, 247, 369, 383, 387, 468, 478–479, 484–487

- power (*cont'd*)
 - to do otherwise, 108–113, 125, 144, 148–149, 382
- possibility,
 - epistemic, 421, 429
 - logical (LP), 47–48, 50, 52–54, 169
 - impermissible/obligation, 97
 - metaphysical (MP, actual and not possible world), 48–53
 - metaphysical (possible world), 96, 159, 163, 207, 247
 - quantum, 323
 - see also* possible worlds
- pragmatism (and free will/responsibility), 11–12, 44, 49, 54–56, 192–193, 195, 198–200, 212, 376, 404, 493–495
- praise, 10, 109, 147–148, 174, 181, 279, 295, 301, 310, 348, 369, 374, 376, 394, 418, 465, 470–471, 493
- praiseworthiness, 10, 95, 100, 109, 147–148, 174, 251, 279, 281, 295, 301, 309–310, 346, 348, 352, 356, 361–363, 369, 374, 376, 380, 384–386, 394, 418, 437–438, 470, 492
 - see also* blameworthiness
- predestination, 112
 - predeterminism, 6, 14
- preference, 186, 351, 393
 - God's, 175, 182
- present,
 - tense states of affairs, 24
 - principle of fixity of, 26, 28, 31
 - see also* luck (present)
- principle of alternative possibilities (PAP), 23, 95–96, 98–99, 101, 125, 210, 269, 294, 386–387, 442, 464
 - see also* alternative possibilities
- probability, 51, 81, 88–89, 181–182, 427, 435
- problem of determinism, 2, 4–6, 10–12, 14–15, 58–62, 64, 72–76
- problem of free will, 1–2, 5–11, 13–15, 59–60, 64, 209, 295, 378–380
- problem of (relation between) free will and determinism, 2–11, 14–15, 58–62, 64, 72–73, 135–136, 235, 281, 370, 376, 481
 - compatibility problem/question, 3–4, 6, 15, 58, 62, 74, 128, 185–187, 195, 197–199, 289–290, 295, 296, 417, 419, 427
 - correlation problem, 60
 - explanation problem, 60–61
 - see also* compossibilism, compatibilism, incompatibilism, impossibilism, moral luck (causal)
- problem of enhanced control, 75, 84–90, 92, 93
 - see also* problem of (relation between) free will and indeterminism
- problem of evil, 110–118
- problem of (relation between) free will and indeterminism, 6, 84–94, 126, 136, 314–334, 369–370
 - see also* libertarian accounts of free will, luck (present), problem of wild coincidences, *Mind* Argument
- problem of wild coincidences, 88–89, 138, 468
- propositions, 14–15, 26, 29–30, 32–36, 71–72, 112, 126–128, 130–135, 139, 141, 227
 - about the past, 126, 128, 131, 135, 137–138
 - about the past and the laws, 126, 131
 - about the remote past, 128–130, 138–139
- psychology, 8, 207, 209, 244, 262, 270, 326–327, 338, 341, 344, 348, 353, 356
- punishment, 8, 10, 64, 113, 189–194, 198–199, 294–296, 325, 376, 398–399, 408, 412, 418, 456, 465, 493
 - see also* finishment, incarceration
- quality of will, 343, 359–360, 366, 385, 396
- quantum physics (and theory), *see* physics
- randomness, 135, 179, 314–316, 321, 369
- rationality, 118, 322, 370, 466
 - see also* irrationality
- reactive attitudes, 114, 193–196, 198, 294, 296, 343, 375–376, 396, 470–472
 - see also* anger, attitudes, blame, hope, love, praise, resentment
- reasons-responsive(ness), 53, 145, 249, 250, 257, 260–261, 263, 270, 338, 342, 385, 440
 - mechanism, 248, 250, 252, 261, 263, 266, 342
 - mechanism ownership, 40, 44, 53, 76
- reasoning,
 - practical, 109, 162
 - regulative control, 15, 485
 - see also* control
- reflectivism, 211
- relationships (personal), 195, 200, 311, 337, 361, 369–370, 375–376, 470–472
- relevance (irrelevance) relation, 5, 14, 15, 68–73, 77, 78, 79, 80, 140, 316
 - see also* incompatibility relation, problem of determinism (explanation problem)
- replacementism, *see* revisionism

- resentment, 193, 196–197, 369–372,
374–376, 398, 402, 471–472
see also reactive attitudes
- responsibility, *see* criminal responsibility, moral responsibility
- revisionism, 8, 11, 198, 204–219, 376, 394,
400, 456, 458
connotational, 210, 213–215, 456, 458
denotational, 212, 213–214, 217, 455, 456
discretionism, 212–213, 217
replacementism, 212, 214
see also eliminativism, compatibilism (revisionist)
- R-indeterminism, 50–52
- sanity, 145, 152, 376, 406, 408, 410–412, 492
see also insanity
- self-control, 146, 174–175, 186, 340, 343,
410–411, 483
- self-defeating problem, *see* time (travel)
- self-forming actions (SFAs), 63, 320
- semi-compatibilism, *see* compatibilism
- skepticism, 11–12, 15, 42, 54, 96–97, 108,
125–127, 134, 138–140, 184, 189, 191,
196–197, 199, 281, 285, 369–377, 455,
464–478
epistemic skepticism, 11–12, 15, 191
free will skepticism (or skepticism about
free will), 11–12, 108, 125–127, 134,
138–140, 184, 187, 189, 191, 198–199,
267–281, 369–376, 418, 430, 464–472,
477–478, 487, 491–492 11–12, 15, 108,
125–127, 134, 138–140, 184, 189, 199,
281, 369–377, 455, 464–476, 487–478
Humean skepticism, 42, 52, 54
luck argument, 378, 380–381, 386
metaphysical epistemic depravity, 493, 495
prohibition, 96–97
skepticism about moral responsibility,
191, 381–383
see also denialism, eliminativism, impossibilism
- skills,
cognitive, 114–115, 224, 346–347, 349
- sourcehood, 4–5, 12, 23, 62–67, 72, 88, 93,
108–111, 125–126, 140, 314, 338, 417,
435, 438
causal, 13, 72, 93, 102, 109, 148, 209, 387,
437–441, 446
constitutive, 5, 8, 13
hard (source) incompatibilism, 4–5, 13
incompatibilist, 7, 111–113, 118, 126, 210
impossibilist, 62, 72
source condition on free will/moral responsi-
bility, 10, 110, 148
source *versus* ability to do otherwise
(classical/leeway), 4, 12–13, 23, 36,
125–126, 140, 338
source view, 126
ultimate, 65–67, 88, 102–103, 118, 148,
172
spacetime, 138, 170
diagrams, 49–50, 52
Spinoza, Baruch, 176, 195
stalemate (dialectical), 5, 492, 494
state of affairs, 24
Stoics, 2, 59, 109, 472
Strawson, Galen, 8, 10, 13, 17, 62, 65–67, 70,
76, 186, 195–197, 205, 210, 380–382,
403, 415, 431, 451, 456–460, 477–478,
488
Strawson, P. F., 11, 196, 207, 216, 243,
375–376, 402–403, 471–477
strong/weak,
ability (or alternatives, can), 101, 104–105,
238
actualization (of worlds), 177–182
case for insanity, 411
causal-dependence, 261, 268, 270
causal invariance, 437
first-order motivations, 174–175
laws, 65
sense of ‘render’, 129, 131, 133–134
views of ignorance, 357
view of intuitions, 295
substance,
agents as, 85, 88–90, 369, 468
temptation, 341
theism, 177, 373
open, 34–36, 112
monotheism/Abrahamic, 169, 180
theodicy, 111, 113
time,
time’s arrow, 231, 233
travel, 157–165
torn decisions, 319–321, 323–324, 326–327,
329
traditional determinism, 2, 12–15
traditional indeterminism, 2, 14
transfer principles, 25–27, 31, 35, 128–131,
133, 135, 137–139, 149, 151–152
 β -like principles, 25–28, 129–135, 137,
141, 294
causal history principle (CHP), 164–165

INDEX

- transfer principles (*cont'd*)
 - non-responsibility transfer, 149, 151–152
 - transfer of the fixity of the past (or present), 28–31, 35–36, 164, 171
- Trinity, 118–119
- truthmakers, 170–171
- trying (to act/choose), 169, 185, 224, 247, 267–268, 270–271, 258–259, 285, 350, 380
- unavoidability, 53, 97, 138, 160
 - of ignorance, 355
 - relations, 138
- ultimate responsibility, 10, 102, 181
- ultimate sourcehood, 65, 67, 87–88, 102–103, 109, 118, 148, 172, 186, 195, 470
 - see also* sourcehood
- unalterability, 97–98, 133–139
 - unalterability version of the consequence argument, 133–139
- unconscious(ness), 9, 210, 243, 246, 249, 251, 263, 265–267, 270, 280–287, 320, 327–330, 407, 477–478
- up-to-us(ness), 125, 380
- urge(s), 266, 280, 305
- utilitarian(ism), 189–191
- van Inwagen, Peter, 2–3, 6–7, 10, 25, 30, 32, 36, 41–43, 58–59, 61–62, 65, 125–140, 149, 151, 294, 369, 378, 387, 419, 434, 442
- values, 116, 172, 175–176, 181, 189, 211, 247–248, 257, 260, 262, 263, 267, 300, 303, 347–348, 361, 364, 381, 427, 438, 481–483
- vetoing, *see* Libet experiments
- virtues, 113, 364, 381
- volitional (states), 267–268, 288, 357, 359
 - prong of law, 410–411
- voluntarism, 176–177
 - versus* rationalism, 177
 - see also* God's freedom, divine leeway
- voluntary (action), 109–110, 114, 116, 173–174, 281, 406, 471
 - involuntary intoxication, 408, 410
- wanton(ness), 341
- will, 109, 280, 340
 - quality of, 343, 359–360, 366, 385, 396
 - weakness of (akrasia), 359
 - will-power, 110, 411
- wrongdoing (-doers), 191, 194, 366, 371–374, 383, 393, 397, 470, 472
- X-Phi, *see* experimental philosophy
- Zygote Argument (Diana/Ernie), 7, 14, 145–149, 435–439, 448, 483–484, 488
 - see also* manipulation (multiple-case) argument
- zombies (free will), 491–495