May Contain Lies

How Stories, Statistics, and Studies Exploit Our Biases – And What We Can Do about It

ALEX EDMANS

UNIVERSITY OF CALIFORNIA PRESS

May Contain Lies

May Contain Lies

How Stories, Statistics, and Studies Exploit Our Biases – And What We Can Do about It

ALEX EDMANS

UNIVERSITY OF CALIFORNIA PRESS

University of California Press
Oakland, California

First published in the UK by Penguin Business, an imprint of Penguin General.
Penguin General is part of the Penguin Random House group of companies.

Library of Congress Control Number: 2024933460

ISBN 978-0-520-40585-1 (cloth : alk. paper)
ISBN 978-0-520-40395-6 (ebook)

Manufactured in the United States of America

33 32 31 30 29 28 27 26 25 24
10 9 8 7 6 5 4 3 2 1

publication supported by a grant from
The Community Foundation for Greater New Haven
as part of the **Urban Haven Project**

Contents

PART I: THE BIASES

PART II: THE PROBLEMS

Introduction

The sweat was dripping down my face as I awaited my grilling in the UK House of Commons. I'd been summoned to testify in front of the Select Committee on Business.[1] This was a group of MPs who, infuriated by a couple of high-profile scandals, had launched an inquiry into how companies were being run.

In my day job as a finance professor, I'm used to being interrogated by students in lectures, journalists in interviews and executives in workshops. But being probed by MPs on live TV and having your testimony transcribed as public record is another level, so I was feeling pretty nervous. I got to the House of Commons early and sat in on the session before mine, burying my head in my notes to swot up on every question the Committee might ask.

My ears pricked up when a witness in that session mentioned some research which sounded noteworthy.[2] It apparently found that companies are more successful when there's a smaller gap between the pay of the CEO and the pay of the average worker. I was intrigued, because my own research shows that employee-friendly firms outperform their peers.[3] My studies don't focus on pay, but this new evidence appeared to complement my findings. For many years I'd been trying to convince companies of the importance of treating workers fairly, and this looked like another arrow to add to my quiver. I wanted it to be true.

If my twenty years in research have taught me anything, however, it's not to accept claims at face value. I pulled up the witness's written statement and saw they were referring to a report by Faleye, Reis and Venkateswaran. But when I looked it up, it seemed to say the exact opposite: the *higher* the gap between CEO and worker salaries, the better the company's performance.

I was confused. Perhaps my nerves led me to misunderstand the study? After all, academic papers aren't known for their clarity. Yet their conclusion was right there on the front page and as clear as day: companies do better if they have *greater* pay gaps.

It then dawned on me what had happened. The witness statement actually quoted a half-finished draft by Faleye, Reis and Venkateswaran that was released three years before the final version.[4] I was looking at the published article, after it had gone through peer review and corrected its mistakes – leading to a completely opposite result.[5]

The witness in question was from the Trades Union Congress (TUC), which holds a strong public position against pay gaps. In 2014, it published a report declaring that 'high pay differentials damage employee morale, are detrimental to firm performance [and] contribute to inequality across the economy’. So the TUC may have jumped on this preliminary draft, without checking whether a completed version was available, because it showed exactly what it wanted.

My own session went smoothly. One question had me stumped, but I told the MPs that I wasn’t an expert in that topic rather than trying to make up an answer. They seemed surprised, as if no one had ever admitted to not knowing something before. In the corridor afterwards, I told the Clerk to the Select Committee about the tainted evidence in the earlier session. He seemed appalled and asked me to submit a formal memo high-lighting the error. I did so, and the Committee published it.

Yet the Committee’s final report on the inquiry referred to the overturned study as if it were gospel. It said: ‘The TUC states that “There is clear academic evidence that high wage disparities within companies harm productivity and company performance” ’ – even though this statement was contradicted by the very researchers the TUC quoted in support. Partly due to this claim, the report recommended that every large UK company disclose its pay gap, and this eventually became law.[6]

The takeaway I’d like to draw is nothing to do with pay gaps – whether they should be published, or whether large gaps are good or bad. Even if bigger differences lead to better performance, we might care about equality more than profits. Instead, it’s to stress how careful we need to be with evidence.

This episode taught me two lessons. First, you can rustle up a report to support almost any opinion you want, even if it’s deeply flawed and has subsequently been debunked. A topical issue attracts dozens of studies, so you can take your pick. Phrases like ‘Research shows that . . .’, ‘A study finds that . . .’, or ‘There is clear academic evidence that . . .’ are commonly bandied around as proof, but they’re often meaningless.

Second, sources we consider reliable, such as a government report, may still be untrustworthy. *Any* report – by policy-makers, consultancies, and even academics like me – is written by humans, and humans have their biases. The Committee may have already felt that pay was too high and needed to be reined in, which is why they launched the inquiry in the first place.

This isn’t just an isolated case. Newspapers publish articles highlighting the blockbuster findings of a study that doesn’t even exist. Companies release research that has no actual data behind it; it just assumes its results. Universities circulate reports declaring game-changing conclusions, when their tests in fact found nothing. Yet if readers want these claims to be true, they accept them unquestioningly.

The problem extends far beyond business. Misinformation surrounds us and affects our everyday lives – how we vote, learn a skill or improve our health. In the

2016 Brexit referendum, buses paraded the claim that European Union membership cost the UK £350 million per week. The actual figure was £250 million, or £120 million after deducting the amount the EU gives back to the UK.[7] People believe the '10,000 hours rule' that you can master any skill with 10,000 hours of practice. Yet the research it's based on was limited to violinists, didn't measure their skill, and didn't even mention 10,000 hours. In 1988, the journal *Nature* published a paper touting the effectiveness of homeopathy, a treatment using heavily diluted substances that supposedly transfer their properties to water.[8] But several other studies found no improvements, and scientific consensus is now that homeopathy is ineffective for any disease or condition.[9]

These examples show how we're all affected by research, even if we never read a single academic paper. Each time we pick up a self-help book, browse through the latest *Men's Fitness*, *Women's Health* or *Runner's World*, or open an article shared on LinkedIn, X or Facebook, we're reading about research. Whenever we listen to an expert's opinion on whether to invest in crypto, how to teach our kids to read, or why inflation is so high, we're hearing about research. And information is far broader than research – our news feeds are bombarded not only with 'New study finds that . . .' but also anecdotes like 'How daily journalling boosted my mental health', hunches such as 'Five tips to ace your job interview', and speculation like 'Why we'll colonize Mars by 2050'.[10] Blindly following this advice, you could find yourself sicker, poorer and unemployed.

In some cases, misinformation can be fatal. In March 2020, as the coronavirus pandemic was breaking out, US President Donald Trump tweeted that hydroxychloroquine might be a cure, proclaiming it 'one of the biggest game changers in the history of medicine'. One woman noticed 'chloroquine' on the label of her fish-tank cleaner; as she told NBC News, 'I saw it sitting on the back shelf and thought "Hey, isn't that the stuff they're talking about on TV?" '[11] She and her husband drank it, hoping it would protect them from the virus. The woman became violently sick but vomited up enough of the chemical to survive. Her husband wasn't so lucky and died just after getting to hospital.

What's striking in all the above cases is that the solution is simple – to check the facts. It seems obvious to ensure a drug is safe before swallowing it, to verify a study exists before writing about it, and to doubt the side of a bus as a source of information. And the people making the misjudgements are more than capable of checking the facts. If I share a study on LinkedIn whose findings people don't like, there's no shortage of comments from executives, investors and fellow academics pointing out how it might be flawed – exactly the kind of discerning engagement I'm hoping to prompt. But do I see the same critical thinking when I post a paper that finds their favour? Unfortunately not: they lap it up uncritically.

One of my favourite toys growing up was Action Man. This UK character was based on the GI Joe set of military figures in the US, which were accompanied by a cartoon series. Each episode closed with a scene where a GI Joe figure taught kids a lesson – don't give your address to strangers, don't pet unfamiliar animals, do wear sun protection. The children in the cartoon exclaimed, 'Now I know!', to which the GI Joe replied, 'And knowing is half the battle.' This aimed to highlight the power of knowledge – with it, you're already halfway there.

But there's another way to interpret that statement: the glass is half empty, not just half full. Even with knowledge, you've *only* won half the battle.[12] Knowing how to check the facts isn't enough. The people who made the above mistakes knew what to do in the cold light of day, yet their biases took over and prevented them applying their knowledge.

As a university academic for two decades, I've seen first-hand how important rigour is when producing research. At the Massachusetts Institute of Technology, where I did my Ph.D.; Wharton, the business school of the University of Pennsylvania, where I was first a professor; and London Business School, where I now teach, I've been held to gruelling standards in my own work. Journals correctly refused to publish my papers until I'd completely nailed the results, addressed alternative explanations for my findings, and toned down any claims that weren't fully supported by the data. Sometimes it took five years of toil and sweat to get a study above the bar for publication.

This isn't just my experience as a producer of research; it's also what I've seen as a gatekeeper. As the Managing Editor of a leading academic journal, the *Review of Finance*, I've been on the other side for six years. After authors submit a paper for potential publication, I send it to 'peer reviewers' (independent experts) and ask their advice on whether to accept it. I've been gratified by the extreme care with which they scrutinize a manuscript. And I've had to apply the same exacting standards myself, rejecting papers that would be highly influential if taken at their word, because their results just weren't identified precisely enough.

While one foot is firmly in academia, my second is deeply rooted in practice, advising companies, investors and policy-makers based on the findings of research. So I've observed how the painstaking care with which papers are written goes out of the window when they're read and emotion takes over. My main field is sustainable business, a field with strong opinions that polarize across political lines. Those on the left tend to believe that ethical stocks always outperform, so they'll trumpet any study which claims this. Many right-wingers retort that sustainable companies are distracted from the bottom line; some US lawmakers have banned state pension funds from investing in them. Sustainability is also a highly practical topic, so I've seen how academic rigour isn't just an academic concept but affects how CEOs run their companies, investors choose which firms to finance, and policymakers decide what laws to pass.

In 2017 I was invited to give a second TEDx talk. It was a great opportunity to reach a wide audience and my instinct was to use it to share my work – as most professors do, and as I indeed did in my first talk. Then I had a thought: what if, instead of pitching my own research, I spoke up for research in general? The whole mission of TED is to promote 'ideas worth spreading', but this mission is under threat if how far an idea spreads depends on whether people like it rather than whether it's true. And it's not just the TED/TEDx stage: anyone with a newspaper column, social media platform, or YouTube channel can broadcast what they want and claim there's data to support it.

So I spoke about how discerning we must be with evidence – how our biases can lead us to fall for something false or reject something real, and how we should judge a study by its carefulness, not its claims. I was grateful when it was elevated to a mainstage TED talk, 'What to trust in a post-truth world', because I hoped it might move the needle, even slightly, from fiction to fact.

Yet misinformation has arguably become worse. Public discourse is increasingly polarized, with opinions formed on ideology, not evidence. The most pressing issues of our time, such as climate change, inequality and global health, are steeped in falsehoods. In the past, we knew what the reliable sources were, such as a doctor or medical textbook for health advice and an encyclopaedia for general knowledge. Now one half of Americans obtain news ‘often’ or ‘sometimes’ from social media,[13] where false stories spread further, faster and deeper than the truth because they’re more attention-grabbing.[14]

And biases exist even among people who’ve seen the talk and should know better. Some companies invited me to present an extended version to their employees, supposedly to promote critical thinking, only to strike out a couple of ‘inconvenient truths’ from the slide deck – because they didn’t want them to be true.

In today’s post-truth world, it’s more important than ever to separate myth from reality. This book is a practical guide to help you think smarter, sharper and more critically – on topics such as how to run a company and invest your money, how to improve your health and develop good habits, how to feed your child and educate a nation’s children, what drives global warming or the spread of coronavirus, and which policies lawmakers should pass and voters should support. We’ll overturn some widely accepted ideas, and in doing so learn simple ways to spot if a claim is supported by the evidence. We’ll uncover the problems with the case study method that pervades the world’s leading business schools, viral TED talks and bestselling books.

We'll see how we can be fooled even by large-scale data – even if hundreds of datapoints all tell the same story.

But knowledge is *only* half the battle. Having knowledge isn't enough: we need to know *when* to use it and *how* to use it. Why do we leave our learnings at the door and rush to accept a statement at face value? Without highlighting the biases that cause us to forget our knowledge, a book that simply passes on knowledge is incomplete. It's like teaching a first-aider how to perform CPR but not how to spot if someone needs it.

Sun Tzu's *The Art of War* stresses that you should 'know your enemy' before drawing up battle plans. So we'll start in Part I ('The Biases') by learning about our enemy. We'll take a deep dive into two psychological biases – confirmation bias and black-and-white thinking – that are the two biggest culprits in causing us to misinterpret information.

In Part II ('The Problems'), we'll study the consequences of these biases. They lead us to climb the Ladder of Misinference shown below:

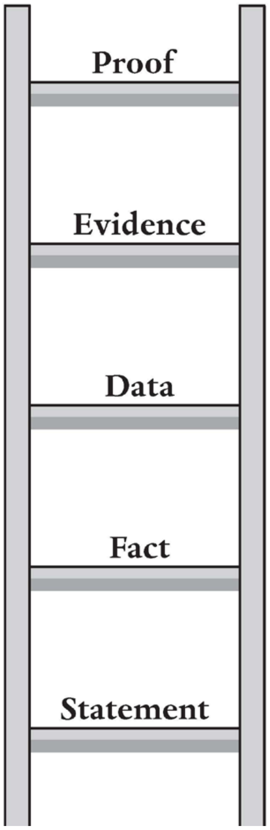


Figure 1. The Ladder of Misinference

We *accept a statement as fact*, even if it's not *accurate* – the information behind it may be unreliable and may even be misquoted in the first place. We *accept a fact as data*, even if it's not *representative* but a hand-picked example – an exception that doesn't prove the rule. We *accept data as evidence*, even if it's not *conclusive* and many other interpretations exist. We *accept evidence as proof*, even if it's not *universal* and doesn't apply in other settings.

Importantly, checking the facts only saves us from the first misstep up the ladder. Even if the facts are correct, we may interpret them erroneously, by over-extrapolating from a single anecdote or ignoring alternative explanations. The word ‘lie’ is typically reserved for an outright falsehood made deliberately, and to accuse someone of lying or call them a liar is a serious allegation. But we need to take a broader view of what a lie can involve so that we can guard against its many manifestations.

‘Lie’ is simply the opposite of ‘truth’. Someone can lie to us by hiding contradictory information, not gathering it in the first place, or drawing invalid conclusions from valid data. The Select Committee’s claim that ‘The TUC states that . . .’ is strictly correct – but it’s still a lie, as it suggests the TUC’s statement was true when the Committee knew it had been debunked. Lies also have many causes – some are wilful and self-interested; others are a careless or accidental result of someone’s biases; and yet more arise from well-intentioned but excessive enthusiasm to further a cause they deem worthy.

This wider definition of ‘lie’ highlights how regulation can’t save us from being deceived – it can only make someone state the facts truthfully; it can’t stop him claiming invalid implications from them. It’s up to us to protect ourselves. Even if a report has been signed off by the government, a paper has been published by a scientific journal or a book has been endorsed by a Nobel Laureate, they should all carry the same health warning: ‘May contain lies’.

Part II thus provides a practical guide to help us discern whether a statement really is fact, a fact truly is data, data genuinely is evidence, and evidence actually is proof. These tips are simple and non-technical, and can be applied even if you're time-pressed and don't have the capacity to dig into the weeds of a study.

To distinguish between truth and lies, and gain a deeper understanding of the world around us, we need to do more than just interpret statements, facts, data and evidence correctly. Part III ('The Solutions') goes beyond the ladder. It moves past evaluating single studies to learning scientific consensus, and assessing other sources of information such as books, newspaper articles, and even our friends and colleagues. From learning how to think critically as individuals, we'll explore how to create smart-thinking organizations that harness our colleagues' diversity of thought, overcome groupthink and embrace challenge. We'll finally examine how to build intelligent societies through teaching critical thinking to our children, taking the politics out of issues such as climate change, and playing our part in the information we share and ignore.

The Appendix provides a checklist of questions to evaluate statements, facts, data and evidence, applying the learnings of Part II. At the start, we might literally go through every question. Over time, the way of thinking that the book develops – challenging what we'd like to believe, listening open-mindedly to what we don't and staying alert to our biases – should become ingrained so we no longer need to follow a script. A novice tennis player thinks, 'First I split-step, then I turn my body so it's square to the net, then I take a backswing and follow through over my opposite shoulder,’ but after a while it becomes second nature.

While this book aims to be practical, it also seeks to be realistic. It’s impossible to overcome our biases in every situation and correctly evaluate every piece of information; the range of ways we can be deceived may seem overwhelming. Our goal is not to become perfect, only better. A baseball player who improves his batting average from 0.280 to 0.320 will leap from a Major League starter to a Hall of Famer, even though he’s still well below 1.000. Critical thinking is a polar star – you might never get there, but it guides you.

Now more than ever, we have easy access to scientific research by the world’s leading minds, yet it’s drowned out by fallacies, fabrications and falsehoods. Knowing what to trust and what to doubt will help us make shrewder decisions, comprehend better how the world works, and spread knowledge rather than unwittingly sharing misinformation. This in turn allows us and our families to lead healthy and fulfilling lives, the businesses we work for and invest in to solve the world’s biggest problems, and the nations we’re citizens of to prosper and thrive. By recognizing our own biases, we can view a contrary perspective as something to learn from rather than fight, build bridges across ideological divides to find common ground, and evolve from simplistic thinking to seeing the world in all its magnificence.

PART I

The Biases

1

Confirmation Bias

Belle Gibson was a happy young Australian. An avid skateboarder with a bright smile, she had the rest of her life to look forward to.

But in 2009, Belle suffered a stroke. Tests to find the cause uncovered devastating news: Belle had an advanced brain tumour and only four months to live. She wouldn’t even see her twenty-first birthday.[1]

Yet Belle was determined. She’d been a fighter her whole life. At just six years old, she had to cook for her brother who had autism, and care for her mother who had multiple sclerosis; her father was out of the picture. So Belle did the only thing she knew: she fought. But chemotherapy and radiotherapy only made her sicker, and after two months there wasn’t the slightest sign of remission.

With conventional treatments totally powerless, Belle’s battling instinct spurred her to fight another way – using natural methods. Weak from the chemotherapy, she forced herself to exercise. She started meditation. She ditched meat for fruit and vegetables.

And, miraculously, she made a complete recovery.

Belle’s story went viral. It was shared, tweeted and blogged about, and reached millions. It showed the benefits of shunning traditional medicine for diet, exercise and sheer grit. Having discovered the secret of how to cure cancer, she now devoted her life to telling others. In August 2013, Belle launched *The Whole Pantry* app, ‘a motivating and supportive resource filled with delicious recipes, wellness guides and lifestyle support’. It hit number one in Apple’s App Store in the first month, as 200,000 people rushed to download the path to better health.

As 2013 drew to a close, *The Whole Pantry* was named Apple’s Best Food and Drink App and was a runner-up for App of the Year. It was so successful that Apple flew Belle to its headquarters to work on a secret project, which, upon arrival, she learned was the Apple Watch – they wanted to develop *The Whole Pantry* into one of its first apps. When the watch was released, Belle’s creation took centre stage, appearing alongside Strava on the Featured Apps page.

Belle wanted to reach those in need through traditional media also. The following year, she published a cookbook, also titled *The Whole Pantry*, with the subtitle ‘Over 80 Original Gluten-, Refined-sugar- and Dairy-free Recipes to Nourish Your Body and Mind’. It was more than just a recipe book; it was a self-help manual that stressed the importance of taking charge of your life. It described how

Belle ‘began a journey of self-education that resulted in her getting back to basics, as she set out to heal herself through nutrition and lifestyle changes’. In her own words, ‘I was empowering myself to save my own life through nutrition, patience, determination and love.’[2]

In just eighteen months, the app and the book netted Belle A$420,000.[3] Yet her motivation was never to make money, but to help others. From day one, she’d pledged to donate a large chunk of her royalties to charity. Due to her success, this amounted to A$300,000.

By 2015, Belle was on the top of the world. She was a best-selling author, a successful businesswoman and a generous philanthropist. Completely cured of cancer, Belle radiated wellness, emboldening others to choose clean eating and good habits over drugs and chemicals in the pursuit of a happier, healthier life.

But Belle’s story was a lie. Belle never had cancer. People believed her without ever checking the facts.

This is a classic example of *confirmation bias*. We accept a claim uncritically if it confirms what we’d like to be true. The public trusted Belle’s account so eagerly, and then spread it so widely, because it struck a chord with our beliefs and values. Belle inspired us to believe you can have anything – even a second chance at life when you’re at death’s door – if you want it hard enough. Ever since we were kids, we’ve been told, ‘You can do anything you set your mind to,’ and Belle was proof.

This bias had serious consequences. Several cancer patients stopped chemotherapy, hoping to emulate Belle. A believer named Kylie explained in a BBC documentary: '[Belle] was saying what she was doing was curing her cancer, it was making it better . . . I had her there to look at, I had her on my phone, she was in magazines, she was on the news, so I trusted her.' Those afflicted by other chronic illnesses, not just cancer, hoped that Belle's formula might work for them also, so they too became disciples.

Sadly, they only became sicker. Belle's followers spurned their doctors' scientifically proven medicines, replacing them with the musings of a blogger. One died within a few months of refusing chemotherapy, prompting her daughter's devastated friend to turn whistleblower to two journalists at *The Age* newspaper in Australia, who broke the story.[4] Hundreds more might have died had Belle not been exposed. After Kylie learned the truth, she restarted chemotherapy and went into remission. The facts saved her life.

We can understand why critically ill patients were easy targets for Belle. If traditional medicine isn't working, you'll seek out alternative cures. But confirmation bias doesn't just prey on the desperation of the sick – an entire cottage industry has sprung up where authors and influencers peddle advice based on emotion, not evidence. Adverts for trading courses promise that we can escape the nine to five and live the life we want, accompanied by the gold standard of social media proof – a photo of the guru cruising in a Ferrari waving a wad of cash. Reality TV stars spread the conspiracy that 5G causes coronavirus, playing into popular mistrust about technology.[5] At the end of 2022, #luckygirlsyndrome amassed 150 million TikTok views within a month. Videos with this hashtag claimed that telling yourself you'll be lucky makes it so – leading some people to blame themselves when things go wrong.

You might think that only the foolish fall for a TikToker's tales. But the temptation to believe alluring claims is so strong that even the rich and famous succumb. Founded in 2003 by Elizabeth Holmes, the medical start-up Theranos claimed it could perform hundreds of tests, including the diagnosis of several life-threatening diseases, from a finger-prick of blood. Investors injected over $700 million, including media mogul Rupert Murdoch, Oracle founder Larry Ellison and former US Secretary of State George Shultz. They'd all reached the top through a lifetime of scrutiny and scepticism – separating the few promising ideas from the hundreds that were pie-in-the-sky. For a small, portable machine to run two hundred simultaneous tests from a few drops of blood was stretching scientific plausibility, so it was essential to put Holmes's claims under the microscope.

Yet many took them at face value, likely because they wanted them to be true. Holmes's story was captivating: a young, charismatic visionary who dropped out of university to pursue her dream and triumph in a male-dominated world. Just as seductive was Theranos's investment thesis – that its shareholders would not only make money but help save lives. One asked for Theranos's audited accounts, wasn't given any, but invested anyway. Another admitted she didn't visit any of Theranos's testing centres or consult any experts, yet still handed over $100 million.[6] In 2015, *TIME* magazine named Holmes one of the world's '100 Most Influential People', and the World Economic Forum inducted her as a Young Global Leader. That October, Theranos was exposed as a sham, and Holmes was convicted of fraud six years later.

Why truth is not enough

You might think the moral of the Belle Gibson and Elizabeth Holmes stories is to always check the facts. But checking the facts is not enough. Confirmation bias is so pernicious that it doesn't just cause us to accept falsehoods. Even if the facts are true, it can lead us to interpret them incorrectly.

Take the case of Bruce Lisker. On the morning of 10 March 1983, seventeen-year-old Bruce swung by the home of his adoptive parents in Sherman Oaks, Los Angeles, to borrow a tool to repair his car. When no one answered the door, he ran to the backyard and looked through the living-room window to see if anyone was in. He glimpsed his sixty-six-year-old mother, Dorka, lying on the floor. Panicking, Bruce darted to the kitchen and removed the panes of glass to climb in, a technique he'd previously mastered to sneak into the house after curfew. He found Dorka unconscious with her head smashed in, her right ear nearly severed, and two steak knives lodged in her back. Bruce pulled them out and called 911. The paramedics came and rushed Dorka to hospital, but she died that afternoon.

The first detective to arrive at the scene was Andrew Monsue, who immediately suspected Bruce. He'd dealt with Bruce before and considered him 'an in-your-face little punk'. Bruce first tried cocaine and LSD aged thirteen, stealing from his parents to support his habit. He had a heated relationship with his mother, which often escalated into screaming matches. In June 1982, he was arrested for throwing a screwdriver at a motorist who'd cut him up and convicted of vandalism.

Monsue instantly decided Bruce was guilty. His theory was that Bruce rifled through Dorka's purse to steal money for drugs. When Dorka caught him, he struck her head with his Little League baseball trophy, followed up with his father's exercise bar, and plunged in the knives. Monsue then interpreted all the evidence he came across as being consistent with this hunch. There were blood spatters on Bruce's shoes and shirt cuffs, which he concluded came from hitting Dorka with a blunt object. Monsue discovered bloody footprints, which, at Bruce's trial, he testified 'resembled quite closely' Bruce's shoes. Soon after Bruce arrived in the county jail to await trial, career criminal

Robert Hughes claimed that Bruce confessed to him. In November 1985, Bruce was convicted of murder and sentenced to sixteen years to life.

Prison was dangerous for a skinny, five-foot-six kid, so Bruce kept to himself, studying computer programming and dabbling in poetry. He wrote a poem about Monsue:

An idiot simpleton who jumped to conclusions;
Unable to reason, 'If not the boy, who then?'[7]

In 2003, after being denied parole multiple times, Bruce filed a complaint against Monsue with the Los Angeles Police Department. Internal Affairs officer Jim Gavin pored through the case files and was surprised to find that Monsue had never ordered a forensic analysis of the footprints, simply going on his hunch that they 'resembled quite closely' Bruce's shoes. Gavin asked forensic scientist Ronald Raquel to examine them. Raquel found prints from two different shoes outside the house, one with a herringbone pattern that couldn't have come from Bruce's tennis shoes. Crucially, this herringbone footprint matched one in the bathroom as well as a bruise on Dorka's head. It also transpired that Hughes had a history of claiming to overhear confessions by other inmates, and had testified against Bruce to reduce his own sentence. In 2009, after twenty-six years in jail for a crime he didn't commit, Bruce was finally freed.

There was no problem with the facts. It's true there was blood on Bruce's shoes and shirt; it's true there were footprints near Dorka's body; it's true that Hughes claimed that Bruce confessed. But truth is not enough. To see why, let's look at one of the most fundamental techniques in statistics. It's called Bayesian inference, and a simplified version is below.

Does ***Information*** support ***Hypothesis?***

Depends on

Is ***Information*** consistent with ***Hypothesis?***

vs

Is ***Information*** consistent with ***Alternative Hypotheses?***

(plus another term)

The *scientific method* involves starting with a *hypothesis* about the world – diet cures cancer, or a suspect is guilty – which we then test by gathering information. A correctly designed test enquires: Does the information *support* the hypothesis? In other words, does the information increase our belief that the hypothesis is true?

But instead we ask: Is the information *consistent with* the hypothesis? 'Support' and 'consistent with' seem like pretty much the same thing – they're even synonyms on Thesaurus. com[8] – but there's a key difference. Even if information is consistent with a hypothesis, it may not support it because of a crucial third question: Is the information consistent with *alternative hypotheses*? Yet, due to confirmation bias, we forget this question and never stop to consider alternative hypotheses, because we're so eager to accept our preferred one.

What matters is not just the facts, but how we interpret them. The blood on Bruce Lisker's shoes was a fact, but Monsue was blind to the alternative hypothesis that it got there from Bruce tending to his mother. The footprints could have been Bruce's, but they could have also been left by someone else. Hughes's claims could have been because Bruce genuinely confessed, or because Hughes was trying to bargain for a shorter sentence. But if we've already formed our conclusion, we interpret any evidence as being consistent with it and it alone.

Bruce is far from an isolated case. US prisoners have spent a total of 30,000 years behind bars for crimes of which they were later exonerated.[9] Criminology professors Kim Rossmo and Joycelyn Pollock investigated what went wrong in fifty of the most serious overturned convictions.[10] There were a number of factors, such as media pressure, unreliable witnesses and flawed forensics, but none of them cropped up in more than half the cases. The clear leader, occurring 74% of the time, was confirmation bias.

We ignore alternative hypotheses in many everyday settings, not just crime. Professors like me would love to believe that education increases income, and stress how people with degrees earn more than those without. However, smarter kids are more likely to go to university, and it could be ability, not education, that boosts their earnings. Motivational speakers reel out testimonials from devotees whose lives were changed after attending their seminars – but those willing to shell out hundreds of dollars and drive for six hours to hear a talk are likely taking other steps to better themselves. Those actions could be causing the turnaround, not the guru's five-point plan.

In all these cases, it's not hard to think of rival theories with a clear head. Yet we don't always have a clear head – we accept our preferred explanation without pausing to consider whether something else could be behind the data. As Francis Bacon, one of the pioneers of the scientific method, noted: 'The human understanding when it has once adopted an opinion . . . draws all things else to support and agree with it.'

Everything we've seen so far is one form of confirmation bias, which we'll call *naïve acceptance* – believing claims we like, without checking the facts or asking if there are alternative explanations. But confirmation bias comes in many different guises.[11] We'll now explore a second.

The danger of denial

20 April 2010 was poised to be a special day for Deepwater Horizon, BP's star drilling rig. Seven months earlier, Deepwater Horizon had dug the deepest well in history, but

it was now embroiled in its toughest challenge yet – creating an oil well in the Macondo reservoir in the Gulf of Mexico. The project was six weeks behind schedule and $58 million over budget, due to weak formations in the ocean rock that required careful drilling.

That day would be a dual celebration. BP executives were on board to celebrate Deepwater Horizon’s stellar safety record of seven years without a single lost-time incident. And the well would finally be complete, netting BP a bounty of 50 million barrels of oil worth $5 billion.

The one remaining step before the rig could be withdrawn was to secure the well. To check it’s safe to do so, you need to run a ‘negative pressure test’. This ensures that the steel casing around the wellbore (the hole you’ve just drilled) can withstand the pressure drop that occurs when the rig is removed and drilling mud[*] is replaced with seawater. You open the top of the drilling pipe, bleed the pressure to zero, close it, and check if pressure builds or fluid leaks into the well.

In the first test, the engineers couldn’t get the pressure below 266 pounds per square inch during the bleed; after closing the pipe, it jumped back up to 1,262. The second attempt did get a zero reading, but it rebounded to 773 afterwards. If that seemed like progress, the third try went backwards – again the pressure went to zero, but then skyrocketed to 1,400. It needed it to stay at zero to pass the test. To put it into perspective, the pressure of a football is approximately 12; for a steam train it's around 250. It wasn't even close.

But 'fail' wasn't the answer the engineers wanted. Deepwater Horizon had an unblemished seven-year record, so in their eyes the test couldn't possibly have been right. Rather than admitting to problems with the well, they decided there must be problems with the test. With huge pressure not to delay the project further, the engineers came up with an alternative explanation for the negative result.

They blamed a 'bladder effect', where heavy mud in the riser (a pipe from the seabed to the water level) exerts pressure on the bladder-like valve that seals the top of the drilling pipe, and the valve then transfers this pressure to the pipe itself. This gave them the excuse to run the test another way – not on the drilling pipe but on the 'kill line', another tube from the seabed to the water surface. They got the zero reading they craved, allowing them to call the test a pass – and sealing the well's fate.

Later that evening, shortly after the executives congratulated the Deepwater Horizon crew on their safety record, gas burst into the casing and travelled up the riser. When it reached the air, it exploded, killing eleven workers and injuring seventeen. Within thirty-six hours, the rig sank. 5 million barrels of oil spilled into the sea, damaging 8 US national parks, endangering 400 species and ruining 1,000 miles of coastline. Local residents and clean-up workers contracted cancer,[12] heart disease[13] and long-term respiratory conditions[14] from breathing in toxic dust and fumes. To this day it remains the worst ever oil spill in the US.

What happened with Deepwater Horizon is a second form of confirmation bias: *blinkered scepticism*. While naïve acceptance involves believing claims that we like and ignoring alternative explanations, blinkered scepticism involves rejecting claims that we dislike and inventing alternative explanations. Such fabrication is known as *motivated reasoning* – grasping at rival theories, no matter how far-fetched, to justify our initial conviction and dismiss the evidence. The US government's official report on the disaster concluded 'there is no such thing as a "bladder effect" that could account for 'pressures the rig crew was observing'.[15] The report's chief counsel put it more bluntly: 'Every industry expert the investigative team met with dismissed the so-called bladder effect as a fiction.'[16]

Blinkered scepticism can apply even if we can't come up with an alternative explanation: we simply dismiss an inconvenient truth without any justification whatsoever. Silicon Valley Bank was the go-to financial institution for many Californian startups and saw its deposits triple between 2019 and 2021. They piled this spare cash into US Treasury Bonds – safe as houses in normal times, but their internal models predicted serious losses if interest rates rose. Rather than heeding this warning, their executives changed the models' assumptions so that they predicted minimal risks. As a former employee told the *Washington Post*, ‘if they see a model they don’t like, they scrap it’.[17]

SVB collapsed in March 2023, the victim of the very interest rate hikes their own models warned about. It was the second largest bank failure in US history, and it was entirely avoidable. Just like Deepwater Horizon, the company saw the iceberg straight ahead – but the bankers put on their blinkers and crashed right into it.

The evidence for confirmation bias

Why do we react so angrily to claims we don’t like? Neuroscientists Jonas Kaplan, Sarah Gimbel and Sam Harris showed how confirmation bias is wired into our brain.[18] They took students with liberal political views and hooked them up to an fMRI[†] scanner. The researchers read out a political statement that the participants previously said they agreed with (such as ‘The death penalty should be abolished’) or a non-political statement (such as ‘The primary purpose of sleep is to rest the body and mind’). They then gave contradictory evidence and measured the students’ brain activity with the scanner. There was no effect when non-political claims were challenged, but countering political positions triggered their amygdala. That’s the same part of the brain that’s activated when a tiger attacks you, inducing a ‘fight-or-flight’ response.

People respond to opposing views as if they're being chased by a wild animal.

The amygdala overrides the prefrontal cortex, the rational part of the brain. In his book *Thinking, Fast and Slow* Nobel Laureate Daniel Kahneman refers to the impulsive, fast thought process (driven by the amygdala) as System 1 and the rational, slow one (operated by the prefrontal cortex) as System 2. In the cold light of day, we know we should read new evidence open-mindedly and try to learn from it – but when our System 1 is in overdrive, the red mist of anger clouds our vision.

Faced with the tiger attack of opposing evidence, we've seen how one coping mechanism is to explain it away. A separate fMRI study investigated what happens to our brain when we do. The researchers recruited 'committed Republicans and Democrats' during the 2004 Bush–Kerry election.[19] They read out a quote from either George W. Bush or John Kerry, followed by a contradictory statement suggesting the politician didn't walk the talk. Finally, they presented an exculpatory statement that justified the inconsistency, similar to what we come up with during motivated reasoning. An example follows:

> *Initial Statement*: During the 1996 campaign, Kerry told a *Boston Globe* reporter that the Social Security system should be overhauled. He said Congress should consider raising the retirement age and means-testing benefits. 'I know it's going to be unpopular,' he said. 'But we have a generational responsibility to fix this problem.'
>
> *Contradictory Statement*: This year, on *Meet the Press*, Kerry pledged that he will never tax or cut benefits to seniors or raise the age for eligibility for Social Security.

> *Exculpatory Statement*: Economic experts now suggest that, in fact, the Social Security system will not run out of money until 2049, not 2020, as they had thought in 1996.

As expected, the contradictory statements lit up the amygdala. What’s particularly interesting is the effect of the exculpatory statements. The researchers found that they activate the striatum, a dopamine-rich area of the brain. Dismissing evidence we don’t like releases dopamine, the same pick-me-up chemical that’s triggered when we go for a run, enjoy a meal or have sex. Motivated reasoning just feels good.

We’ll now move from the origins of confirmation bias to its effects. Stanford psychologists Charles Lord, Lee Ross and Mark Lepper took a group of students who were either highly proor highly anti-capital punishment and gave them a summary of two papers:

- A *time-series* study that compared the murder rate in a single state over time, before and after the death penalty was introduced. For example, ‘Kroner and Phillips (1977) compared murder rates for the year before and the year after adoption of capital punishment in 14 states. In 11 of the states, murder rates were *lower after* adoption of the death penalty.’
- A *cross-sectional* study that compared murder rates across pairs of states, one with the death penalty and

> one without, at a single point in time. For example, 'Palmer and Crandall (1977) compared murder rates in 10 pairs of neighbouring states with different capital punishment laws. In 8 of the 10 pairs, murder rates were *higher* in the states *with* capital punishment.'[20]

The studies were randomly assigned across the students. For some, the time-series analysis supported the death penalty and the cross-sectional one opposed it, as in the above example; for others it was the reverse. Importantly, the studies were fictional, removing concerns that one was genuinely better than the other.

The researchers asked the students to rate the rigour of each report. If the rational System 2 were in charge, their assessment should have been driven entirely by its methodology and not its conclusions. Instead, the subjects criticized the research if it contradicted their view and had no trouble proposing alternative explanations. If the time-series study found that the murder rate rose after capital punishment was introduced, a death-penalty supporter claimed that it would have jumped even faster without the law change. But if a paper found the result people wanted, they accepted it with no questions asked.

The students were then asked whether their beliefs about capital punishment had changed after reading the research. Each had seen one study supporting their stance and one challenging it, so the mixed evidence should have led them to moderate their view. Instead, their initial beliefs strengthened – opponents of the death penalty became even more hostile, and supporters yet more fervent. They'd seized on the paper they liked and ignored the other.

It's worth pausing a moment to highlight the significance of this finding. We might think that disagreement is caused by differences in information, and so it's solved by sharing information. If we lay our facts on the table, the other person will see things our way. The results suggest it's not that simple. Even if two people see the exact same data, their opinions can still differ.

This is known as *belief polarization*. Someone who doesn't like the information will find a reason to ignore it, while a supporter will see gospel in myth. It's like a football match where opposing fans argue furiously about a penalty kick, even though everyone watches the same game. We see what we want to see.

Naïve acceptance and blinkered scepticism are two different forms of confirmation bias, but also two sides of the same coin. They both lead to *biased interpretation* of evidence – we believe what we want and ignore what we don't. However, remember that confirmation bias comes in many guises. The third has nothing to do with how we interpret information, and it's to that we now turn.

Scholar's Mate and chess traps

The first hobby I ever had was chess. I began when I was five and eventually made it to the England junior team, but it’s been decades since I last launched a Fried Liver Attack or fended off an opponent’s Frankenstein–Dracula Gambit. When I started out, I learned all the juiciest traps you can lure an opponent into. The best is Scholar’s Mate, where you get your Queen – your strongest piece – out early, hoping your opponent won’t notice it’s about to give checkmate. If all goes well, you can win in four moves.

Since Scholar’s Mate is so well known, even novices are unlikely to fall for it, so I made up my own traps. After I laid the ambush, I’d pray, ‘Please let her take my pawn’ under my breath, trying to maintain my best poker face. Which, as an excitable five-year-old, wasn’t particularly convincing.

I beat many opponents that way. But then I faced stronger adversaries who’d spot the trap with ease and counter with a response that made my original move look foolish. Without knowing it, I’d fallen for another type of confirmation bias. While *biased interpretation* concerns what we do with information once we get it, *biased search* is about what information we gather in the first place. We’ll only look for evidence that confirms our initial hunch and don’t dare poke around for something that might contradict it.

When thinking about which move to play, I’d consider all the possible ways my opponent would fall into the trap, justifying my candidate move as the right one. Instead, I needed to investigate the ways she could *refute* my move. Only if there was no way of doing so was it safe to play it.

This problem applies far beyond the chessboard. Psychologist Peter Wason uncovered the first systematic evidence for biased search with the following study.[21] He gave people a set of three numbers and asked them to identify the rule that generated them. They tested their guess by proposing a separate trio, and Peter would say whether this set followed the rule.

Let's take an example. If you saw the numbers 2–4–6, what's the rule? Most of Peter's subjects thought 'successive even numbers'. They tested it by proposing other sets of successive even numbers – perhaps 4–6–8, 12–14–16, or 218–220–222. And Peter would say 'yes', consistent with the hypothesis.

But the 'yes' doesn't *support* the hypothesis. Knowing that these triplets also satisfy the rule tells you almost nothing, because it's also consistent with *alternative hypotheses*. Perhaps the rule is any three even numbers, or any three increasing numbers – so this new information is nearly useless.

The only way to support your theory is to try to disprove it, just like trying to refute your own move in chess. You might test something like 4–12–26. Getting a 'yes' would rebut your theory of successive even numbers, indicating that you need to explore rival theories. And if you heard 'no', that would reject 'any three even numbers' and 'any three increasing numbers’ – by ruling out the alternative hypotheses, it supports yours.

Yet most people won’t try the 4–12–26, because they won’t entertain the possibility they might be wrong. They’re worried about getting a ‘yes’, refuting their hunch, and this fear causes the amygdala to light up. But finding out what’s wrong is the only way to find out what’s right. Lightbulb inventor Thomas Edison is quoted as saying: ‘I have not failed. I’ve found 10,000 ways that won’t work.’

You might think it doesn’t really matter what people think about a series of numbers. However, biased search can also blind us on core beliefs that affect our health, and even whether we devote our lives to serve a god. Timothy Brock and Joe Balloun asked 112 undergraduates to listen to recordings of speeches by high-school students and judge how persuasive they were.[22] The university students thought the researchers were interested in their ratings to provide feedback to the schoolkids, but they in fact had another goal.

There was static in the recordings, but Timothy and Joe told them they could press a button to get rid of it. One speech argued that Christianity is evil, another that smoking didn’t cause cancer, and a third that it did, in addition to three neutral talks used as benchmarks. After the students rated the talks, they filled in a survey which asked how often they went to church and how much they smoked, alongside neutral questions.

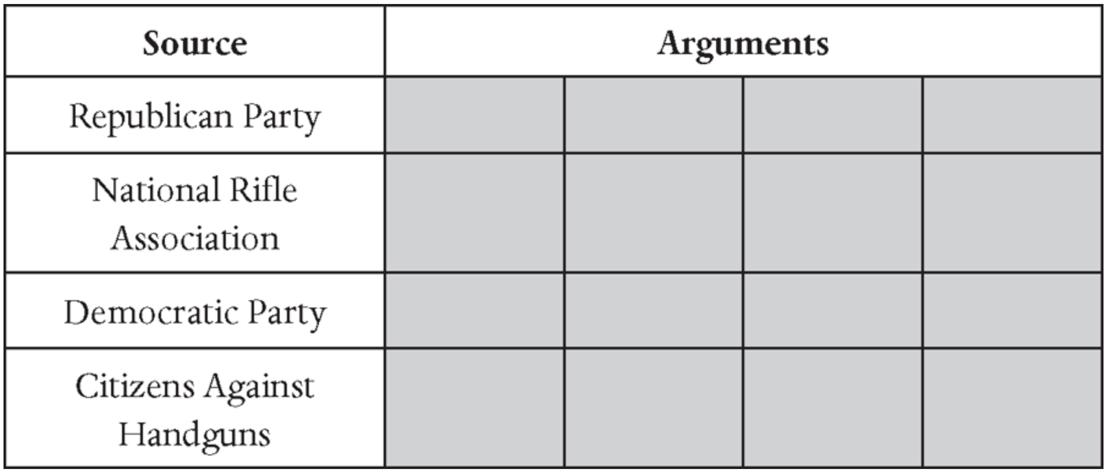
The experimenters found that students only got rid of the static if they were disposed to like the message. Churchgoers didn’t want to hear the case against Christianity and were happy that the recording was garbled. Smokers closed their ears to the dangers of lighting up but wanted clarity for a speech refuting anti-smoking arguments.

Biased search means that the huge rise in the availability of information may actually make us less informed. Nowadays, we can look something up on our phone instead of trekking to the library; scientific research is increasingly ‘open access’ rather than hidden behind a paywall. This should lead us to become more balanced, as it’s easy to learn about both sides of an issue. However, the opposite is true – it increases the likelihood we can corral some evidence that supports our viewpoint, and so beliefs become even more binary.[‡] Biased interpretation means that we see what we want to see; biased search means that we find what we want to find.

How knowledge backfires

We’ve established that the average citizen suffers from confirmation bias. But surely knowledge is a cure? More knowledgeable people might better appreciate the logic in a counterargument and seek out different views to further broaden their understanding. If so, confirmation bias might not be that much of a problem. Sure, it causes the person on the street to make mistakes, but we’d hope the politicians who govern our country, the executives who run our company and the investors who manage our pension are smarter than us mere mortals. Or, some readers might think their intelligence will protect them from confirmation bias so they’d never make the mistakes discussed in this chapter.

Sadly, this isn’t the case. Charles Taber and Milton Lodge found that knowledge makes things worse.[23] They first explored biased search by asking political science undergraduates to research gun control and, importantly, in an even-handed way so that they could explain the debate to others. The students had access to four information sources, and each source had four arguments, as in the table below:



Source	Arguments			
Republican Party				
National Rifle Association				
Democratic Party				
Citizens Against Handguns				

Clicking on one of the grey boxes would reveal an argument, but they could only click on eight out of the sixteen, so they had to be selective. Students with low political knowledge searched in an even-handed way – on average, they chose only slightly more than four arguments that supported their own views on gun control and slightly fewer than four against. Their more informed peers showed a sharp imbalance, choosing six consistent opinions and only two contrary ones. A firearms opponent was eager to hear what Citizens Against Handguns had to say but didn't care for the views of the Republican Party; the reverse was true for gun supporters. Knowledge doesn't make you more aware of the need to consider both sides; instead, you engage even more in biased search.

The second part of the experiment turned to biased interpretation. The subjects were now given four views on affirmative action, two on each side, and asked to comment on how convincing they were.[§] For arguments that supported their own stance on affirmative action, less informed students gave an average of two favourable and one unfavourable remark (2–1); for opposing arguments it was 1–2. The differences were much starker for more informed students: they gave 3–0 for views they liked but 0–6 for those they disliked. Rather than allowing you to see both sides of an issue, knowledge helps you conjure up more reasons to praise opinions you share, and poke more holes in those you resent.

In a nutshell

- Confirmation bias leads to *biased interpretation*. This comes in two forms.
  - *Naïve acceptance* : believing claims that we like, without checking the facts or asking if there are alternative explanations.
  - *Blinkered scepticism* : rejecting claims that we dislike, and inventing alternative explanations – known as *motivated reasoning*.
- To detect confirmation bias, ask: Do I want this statement to be true?
  - If yes, be mindful of naïve acceptance and ask if there are rival theories.
  - If no, be mindful of blinkered scepticism and take the claim seriously.
- Confirmation bias is hardwired within us. Statements we dislike trigger our amygdala, the part of the brain that induces a fight-or-flight response. We get a dopamine hit if we're able to dismiss them.
- Two people can look at the same data yet draw different conclusions; they see what they want to see. Putting the facts on the table may lead to *belief polarization*, where people's views become more opposed, not less.
- Confirmation bias also leads to *biased search* – we close our eyes to sources we'll disagree with and find what

we want to find. Yet the best way to support a hypothesis is to try to contradict it.

- Knowledge doesn’t make us more aware of our biases. Instead, it can make us more susceptible to them, by allowing us to engage in motivated reasoning. It also worsens the problem of biased search.

Even though confirmation bias is widespread, it doesn’t apply to every case in which we encounter information. Emotions run high for the death penalty, gun control and religion, but for many day-to-day decisions we don’t have a prior view. If there’s nothing to confirm, there’s no confirmation bias, so we’d hope we can approach these issues with a clear head.

Unfortunately, another bias kicks in. It’s to this bias that we now turn.

2

Black-and-White Thinking

In 1947, Ohio teenager Robert Coleman had a bright future ahead of him. The son of a candy salesman and a homemaker, he'd just placed second out of 8,500 seniors[*] in a statewide scholarship test and been accepted to the University of Michigan's pre-med programme.

Robert wasn't just book smart. He soon became an icon on the Michigan campus for his jokes and impersonations, which he'd crack out at parties. After graduating in 1951, Robert spent the summer as a comic waiter and entertainer in New York State's Catskill Mountains. He was such a hit that a talent scout offered him a contract. Robert was on the brink of signing when he mentioned – in the same off hand manner he delivered his punchlines – that he'd originally planned to become a doctor. The scout, knowing all too well the unstable life of a comedian, ripped up the contract and urged him to return to medicine.[1]

So Robert enrolled at the prestigious Cornell University Medical School for his MD, and then completed a residency in cardiology in 1959. Most young doctors begin their careers working for a hospital, but Robert wanted to be his own boss. Straight out of residency, he opened his own practice on the wealthy Upper East Side in New York City. It got off to a slow start, causing Robert, a self-confessed food lover, to become depressed and overweight. He'd ballooned 80 pounds since graduating from high school, and now tipped the scales at 225.

Robert went back to the medical journals he'd pored over as a student, desperately searching for a diet that might help him shed weight and, importantly, do so without him feeling hungry. One study, 'A new concept in the treatment of obesity', sounded promising.[2] He took the plunge and built a programme based on its findings. It was an instant hit – Robert lost 30 pounds in four weeks and didn't suffer any hunger pangs.

This success inspired Robert to make a radical career shift from cardiology to weight loss. The sluggish progress of his private practice turned out to be a blessing in disguise, as it had forced him to take side jobs with companies, monitoring their executives' health. These connections gave him the ideal lab to test his diet on a larger scale. He put sixty-five executives from telephone company AT&T on his programme, and all but one reached their target weight.

Robert's fame spread. In 1965, he gained national recognition when he promoted his weight-loss plan on *The Tonight Show*. *Vogue* magazine published his diet five years later; it became so popular that it was dubbed 'the *Vogue* diet'. A book deal followed in 1972, and Robert Coleman sold a million copies in just four months. The book eventually spent five years on the *New York Times* bestseller list and was bought by 15 million people.

But Coleman was only Robert's middle name. His full name was Robert Coleman Atkins, and the *Vogue* diet later became known as the Atkins diet. His book *Dr Atkins' Diet Revolution* contained no references, footnotes, or even a bibliography. The publisher had ordered Atkins to delete them, warning that too much science would diminish the book's appeal: 'This isn't a medical book, it's a popular book that will be bought by people who don't normally read books.'

And true enough, millions around the world combed through the book and overhauled their eating habits based on no more than anecdotes. The diet they were following had no scientific evidence to back it up – not even on whether it achieved its claims of short-term weight loss, and certainly not on any long-term side effects. Yet the book that promoted it is the bestselling weight-loss book in history.[3]

The Atkins appeal

So why was the diet so popular? Because it was simple. It had one rule, and only one rule: Avoid all carbs.[†] Not just refined sugar, not just simple carbs, but all carbs. You can decide whether to eat something by looking at the 'Carbohydrate' line on the nutrition label, without worrying whether the carbs are complex or simple, natural or processed.

The Atkins diet didn't consider whether carbs could be good in moderation – perhaps they're OK so long as they don't exceed 50% of your daily calories? That rule would be more forgiving and flexible than 'minimize carbs'. However, it would be hopelessly complicated to follow. You'd have to measure the total carbs you consume each day, and also your protein and fat. You'd then convert each nutrient into calories, taking into account their different calories per gram. It'd be a huge hassle that would never catch on. Instead, Atkins's dictum was simple – virtually no carbs.

The Atkins diet played into *black-and-white thinking*. This bias means that we view the world in binary terms. We see something – be it something concrete like carbs or red wine, something abstract like religion or capitalism, or something practised like weight-training or meditation, as either always good or always bad. In reality, maybe it has no effect at all, it's good or bad in moderation, or some types are good but others are bad. But there was no such nuance with the Atkins diet. Carbs are sinful; fat and protein are saintly. Full stop.

In Chapter 1, we explored confirmation bias – the temptation to accept a claim uncritically if it supports your beliefs. If you're an aficionado of Amarone, a connoisseur of Chianti, or a dilettante of Dolcetto, you'll latch on to a study declaring that red wine is good for you. Yet black-and-white thinking can be more pervasive than confirmation bias, because it makes you a sucker for one-sided advice regardless of which way it goes, and *even if you don't have a prior view*. You might have no idea whether something is black or white, but if you think it must be one or the other – you're not open to shades of grey – you'll be swayed by unambiguous claims.

Most people think 'protein' is good. You learn in primary school that it builds muscle, repairs cells and strengthens bones. 'Fat' just sounds bad – surely it's called that because it makes you fat? 'Carbs' aren't so clear-cut. Before Atkins, people might not have had strong views on whether they're good or bad. If the Atkins diet had recommended eating as many carbs as possible, it might still have spread like wildfire.

To pen a bestseller, Atkins didn't need to be right. He just needed to be extreme.

Caveman logic

Why are we wired to see the world as black and white?

Let's take a little step back in time – by a mere 1 million years – to the era of our hunter-gatherer ancestors. A hunter-gatherer’s life was tough. To enjoy meat for dinner, you’d first have to make a spear with tips fashioned from flint or bone, and then hunt an animal for up to five hours in the midday sun. If you were particularly skilful, and lucky, you’d kill it, but your work was far from done. You had to skin the animal, build a fire and cook the meat. Even after supper, there was no time to rest, as you made clothes and shelter from the discarded skin before repeating it all the next day.

All this was while ensuring you didn’t become prey yourself to hyenas or tigers, or be killed in competition for a mate. These dangers meant that a hunter-gatherer’s lifespan was just thirty years, and this precarious reality forced him to make snap judgements. What if he’d spent the whole day hunting an antelope, but it eluded him? He’d then need to forage for food and decide what to eat. Some food might be poisonous, but with a starving family there was no time to test it – feed it to an animal, perhaps – he had to think fast. So he followed a black-and-white rule, like ‘berries are good’. He also needed to figure out which animals were predators but could be overcome, which were dangerous and should be evaded, and which weren’t a threat and could be disregarded. Again, he needed a quick decision – fight, flight or ignore? So he’d use a simple formula, such as ‘run from carnivores’.[‡]

Black-and-white thinking allowed hunter-gatherers to act decisively in life-or-death situations where speed was of the essence. However, simple rules occasionally led to mistakes. Some berries were poisonous and caused sickness or even death. Running from all carnivores meant you wasted energy, abandoned a food source or didn’t domesticate a dog that could help with the hunt.

Fast-forward a million years and the world is a different place today, with few life-or-death situations. While speed is less critical, scale has become crucial due to the volume of decisions we make. Black-and-white rules are attractive as they’re easy to learn and can be applied en masse – but following them blindly can sometimes lead us astray. When spelling, it’s ‘i’ before ‘e’ except after ‘c’, yet counterexamples exist, such as ‘seize’. In science, we’re taught that heat causes something to expand and cold makes it contract, but there are exceptions, like water becoming ice. Music theory teaches us the four chords that should be the heart of every song – if you’re in the key of A, these are A, E, F sharp and D.[§] But songs can break the rules by doing something unexpected. The verse of Oasis’s ‘Champagne Supernova’ sits in A. Then when Liam Gallagher switches to the chorus and gets to the word ‘landslide’, our ears prick up as he slides down from A to G, underlining the lyric.

Another cause of binary thinking is that everyday life is arranged into binary divisions. A sports game has two teams, so players with a different-coloured shirt are the enemy. You don’t pass to them; if they have the ball, you try to steal it back. In many choices, we’re given only two options – a typical US Presidential election has just two viable candidates; in the UK's Brexit vote, it was Leave or Remain. As a result, we interpret a piece of information as supporting only one or the other.

Convinced? You shouldn't be . . . at least not yet. The above examples only lay the groundwork for black-and-white thinking – it's in our ancestry, helps us make decisions at scale and results from how the world is divided – but they don't amount to scientific evidence that it exists. For this, we need academic studies on how people process information. There are three different strands of research, each examining a different way in which black-and-white thinking leads us astray. Let's now explore these three errors.

The dangers of black-and-white thinking

The common answer to why black-and-white thinking is flawed is that 'the world is shades of grey'. That's indeed right, but we can be more precise: the world may be moderate, granular or marbled, as shown in Figure 2.

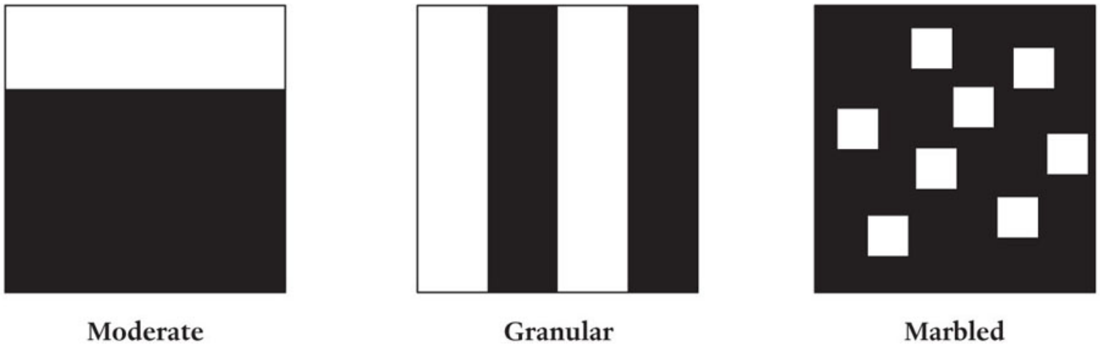


Figure 2

Moderate

Something is *moderate* if it’s bad only after a point; before that point, it’s good. Carbs are an example: based on scientific research, the US National Academy of Medicine recommends that they comprise 45–65% of your daily calories. Yet the Atkins diet calls for near zero.[4]

When the world is moderate, viewing it as black-and-white can be dangerous. A *Lancet Public Health* study found that a fifty-year-old with a carb intake of under 30% has a life expectancy four years shorter than one with 50–55%.[5] Even that underestimates the danger, since the

Atkins diet recommends near zero, not just below 30%. Other research documented kidney problems from the high-protein intake and heart disease due to the saturated-fat consumption.[6]

Atkins himself may have been a casualty. He suffered a heart attack in April 2002 and was hospitalized for a week. He eventually recovered but slipped on an icy pavement the following year, hitting his head and going into a coma; he died nine days later. A copy of the medical examiner's report into Atkins's death was later leaked to the media. The report mentioned heart disease, which his opponents attributed to overconsumption of animal proteins and saturated fat.

The Atkins diet demonizes carbs as always bad, when moderation means they're only harmful after a point. Alternatively, we may view something as always good, when moderation means it's only beneficial up to a point.

On a Sunday morning in April 2007, David Rogers was pumped. The twenty-two-year-old fitness instructor was about to run his first ever marathon – and not just any marathon but the iconic London Marathon. Two years earlier, he'd conquered the Great North Run, the most popular half-marathon in the world, but a marathon is a big step up. Yet David had trained tirelessly throughout the cold English winter and was confident. He wasn't just running for himself either – he'd raised £1,200 for the Motor Neurone Disease Association, and his parents and friends had trekked down to London to support him.

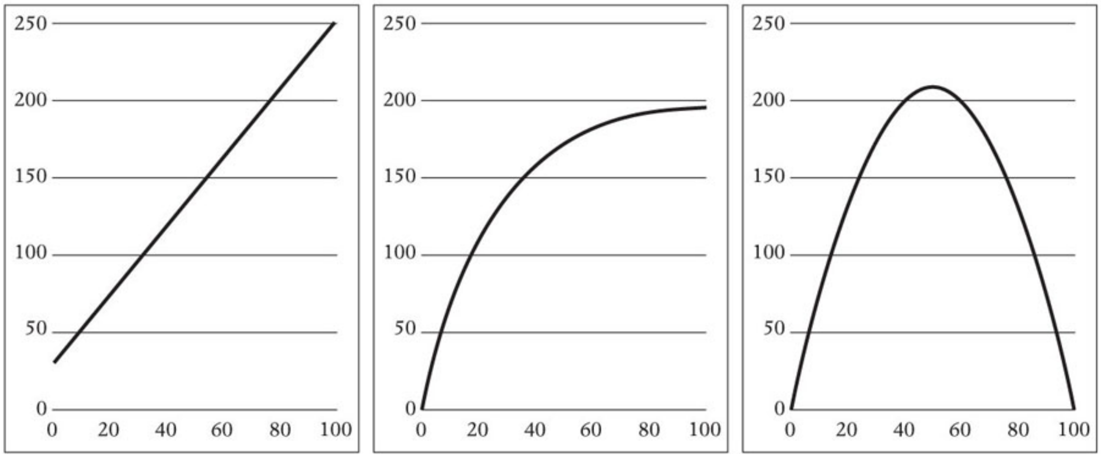
The 2007 London Marathon turned out to be the hottest in history, with temperatures peaking at 23.5°C. When the weather forecast was released, a flurry of articles and blog posts popped up offering advice. Realizing that the runners would have trained in the winter and weren't used to the heat, their biggest tip was to hydrate as much as possible – the night before, the morning of and during the race. Demand for water would be so high that supplies might run out so runners should sip from every station, even if they weren't thirsty – they couldn't be sure where the next watering hole would be.

David dutifully followed that advice. He stayed hydrated throughout and completed the race in 3 hours 50 minutes, beating the 4-hour benchmark that's widely viewed as a good time. Shortly after crossing the finish line, David collapsed and was taken to hospital, where he died from water intoxication – drinking so much water that essential minerals like sodium are diluted to dangerous and, in David's case, fatal levels.

Moderation exists in less extreme situations than life and death. Fitness apps track your exercise minutes as if more is better – it's common sense that lifting more weight builds more muscle. Yet lifting doesn't actually grow muscles, it tears them. It's on the rest days that the muscles repair and grow back stronger. Many business practices peddled by books, case studies and management consultants are only good up to a point. A boss providing feedback to her workers helps them develop, but too much feedback, too often, can make them feel micromanaged. Giving employees free rein can spark innovation, but excessive empowerment may lead to uncoordinated efforts and a lack of direction.

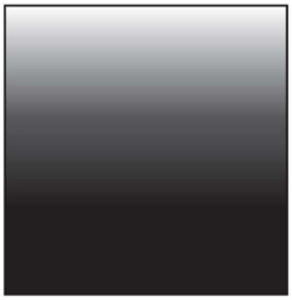
Let’s now turn to the promised evidence. Psychologists Edward DeLosh, Jerome Busemeyer and Mark McDaniel show how we think in black-and-white ways – we view effects as always positive (water always improves marathon performance), or always negative (carbs always hinder weight loss), even when the data clearly shows a moderate relationship.[7] They gave students data on how an unknown substance affects human arousal. They first told them how much a person had ingested, on a scale of 0 to 100, and then asked them to estimate how turned on they were from 0 to 250. After the students had guessed, the actual level of stimulation was shown. This process was repeated 200 times, so there was ample opportunity to learn the true effect.

The 108 students were divided into three groups of 36. For two, the impact of the substance was always positive – the first at a constant rate, the second at a decreasing rate. The third saw moderation – a greater dosage first increased arousal, but then lowered it. These relationships are shown below:



At the start, the students in the third group made the greatest prediction errors, probably because they thought the effect would only go one way. Even after 100 datapoints, the errors were still twice as high as the second group – black-and-white thinking is so ingrained in us that it’s difficult to unlearn, even when the data clearly shows that more sometimes leads to less. Only by the end of the 200 trials did the third group accurately predict the relationship. And then, in a second set of 200 trials, many of them forgot what they’d learned and regressed to predicting an always-positive effect.

The first diagram in Figure 2 is a big improvement over black-and-white thinking, because it contains both colours. Yet it still oversimplifies reality because it contains only black and white, not shades of grey. It implies that there are bright lines that can’t be crossed – water is always good, but suddenly becomes poisonous if you exceed 1 litre per hour; carbs are fine, but woe betide anyone who crosses 65.00% of their daily calories. People interpret the 2015 Paris Agreement as suggesting that a 1.5ºC temperature rise is a tipping point that, if exceeded, would cause the world to end and render future climate action futile. Instead, reality is more like the diagram below, where increasing the amount of something gradually causes more harm but there’s no threshold below which it’s totally safe, nor above which you might as well give up.[‖]



Moderate

That’s not to say targets aren’t useful – they give you something to aim for, and you can hold yourself accountable for whether you reach them. Aspiring to run a marathon in under four hours gives purpose and direction to your training compared to ‘I’ll do my best’, a weekly practice target disciplines you to open your clarinet case three times a week, and the Paris Agreement brought a level of unprecedented global coordination to one of the world’s most serious challenges.

The problem arises when you view hitting the target as all or nothing. If your training times suggest that four hours is out of reach, you might throw in the towel and opt for a Netflix and ice-cream marathon instead. A November 2022 issue of the *Economist* quoted a variety of climate models predicting that 1.5ºC would almost certainly be overshot but stressed that 'every fraction of a degree matters'.[8] If governments excessively fixate on 1.5ºC and view it as unachievable, they may simply wave the white flag and not bother with climate policy. Something is *granular* if it comes in many different forms, some of which are good and others bad. 'A new concept in the treatment of obesity' warned about 'the stimulating effect of glucose on lipogenesis [fat creation] and . . . the poor utilization of glucose for energy purposes by obese subjects'. That statement refers only to glucose, but the Atkins diet extrapolates from it to denigrate all carbs. In fact, complex carbs (like rice, quinoa and potatoes) are much better than simple carbs, because the body takes longer to break them down, so they don't cause a sugar spike.

Granular

Granular

The above example involves taking a study about a single tree (glucose) and generalizing the results to the entire forest (carbs). We also make the opposite mistake of thinking that a statement about the forest applies to every tree. Children learn at primary school that cholesterol clogs your arteries. In fact, high-density lipoprotein cholesterol is good – in moderation, of course – because it cleanses 'bad' cholesterol from your bloodstream.

Scientific research shows how we ignore granularity and instead engage in *categorical thinking*, where we put items into buckets and make decisions based on these buckets, not the individual items. A particularly vivid field that does so is 'disgust research'. Yes, that's a legitimate research area – and a fertile one, with the University of Pennsylvania's Paul Rozin a leading expert.

In one experiment, Paul and two colleagues poured apple juice into a brand-new bedpan and asked people to drink it. Even though they acknowledged the bedpan was perfectly clean, 72% flatly refused to drink from it.[9] That's because our brains have a category for drinking vessels, and toilets aren't in that category.

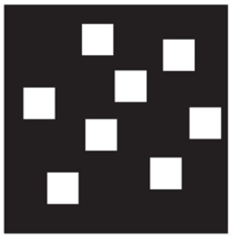
In another study, Paul and different co-authors sterilized a dead cockroach and briefly dipped it into a glass of juice.[10] Out of fifty students, only one drank the juice, yet they were entirely willing to do so if a plastic candleholder was dunked instead. A third experiment found that people refused to eat high-quality chocolate fudge shaped – with remarkable realism – into dog faeces, while they were quite happy to eat square-shaped fudge. Again, categorical thinking is at play. We classify insects as dirty and think they can't possibly be sterile, and items shaped like dog poo as disgusting even if they're delectable.

Paul's studies concern yes/no decisions – you either consume or you don't. Other research shows that categorical thinking also exists for continuous choices, where you select from a range. Joachim Krueger and Russell Clement took 177 undergraduates at Brown University in Providence, a city in the north-east of the US.[11] They were given a date (such as 30 May) and asked to guess the average temperature in Providence on that date. This was repeated for forty-eight different dates, four for each month, presented in random order.

The researchers found that the students' forecasts depended on the month, rather than the actual day within that month. Their estimates for the May dates (2, 9, 16, 23 and 30 May) were all similar to each other, even though May is granular and the 30th is warmer than the 2nd. They then forecast a sudden jump between 30 May and 7 June, even though 30 May is likely much closer in temperature to 7 June than 2 May. It's as if they had in mind 'May is warm, June is hot' – irrespective of the actual date.

Even experts suffer from categorical thinking. On 27 January 1994, Willard Scott, a famous weatherman on NBC's *The Today Show*, exclaimed, 'Jeez, come on, February!' Yet the temperature doesn't suddenly leap when you turn your calendar from 31 January to 1 February.

Marbled



Marbled

Something is *marbled* if it contains both good and bad elements, so you can’t easily classify it as either black or white.

Let’s look at an example. Many investors are concerned about climate change, so they have policies to avoid all fossil-fuel stocks. However, oil and gas firms might not be unambiguously bad. Many are investing urgently in renewable energy because they know they need to reinvent themselves; indeed, the energy sector produces more green patents than almost any other industry.[12] A black-and-white policy of shunning all energy stocks may then backfire, as it denies funding to companies with the most potential to address climate change. Similarly, semiconductor companies are often at the top of lists of globalwarming contributors. That’s because the manufacturing process releases perfluorocarbons, which trap far more heat than CO2. Yet semiconductors are used in solar panels and other renewable energy converters,[#] making them crucial in the fight against climate change.

Note that marbling is subtly different from moderation and granularity. Moderation suggests it's fine to invest £1 million in an energy company, but not £2 million. Granularity implies that some oil and gas stocks are unquestionably good and others are flat-out bad. Marbling suggests that the entire sector is nuanced – it contains both positive and negative elements, just like marbled meat has streaks of fat interspersed with the muscle.

An experiment by Florian Heeb and co-authors found that people evaluate sustainable investments in a black-and-white manner, even if they're marbled.[13] They're willing to pay more for green stocks, but this willingness doesn't depend on how green the investments are – it doesn't matter if they save 5 tons of CO_2 or 0.5. They label investments as green or not green, ignoring the actual level of sustainability.

Psychologists Matthew Fisher and Frank Keil similarly uncovered black-and-white thinking in the way we interpret marbled information.[14] They gave a group of 150 students several pieces of eyewitness testimony. Some were unambiguous, such as 'One witness is 100% confident that the defendant did not commit crime X,' but others were mixed, such as 'One witness is 50% confident that the defendant did not commit crime X.'

After seeing all the evidence, each subject was asked how likely it was that the defendant was guilty. Matthew and Frank found that this assessment was driven by the *number* of proversus anti-statements, not their strength.

50% confidence had as much impact as 100% certainty – marbled evidence was interpreted as black or white.[**] As they concluded, people fail ‘to properly weigh the strength of a given piece of evidence and instead evaluate in an all-or-none manner’. Shades of grey get lost in the shadows if we only look for black and white.

In a nutshell

- *Black-and-white thinking* means that we see something as always good or always bad. Even if we have no prior view on a topic, we’re more likely to believe a statement if it’s extreme or sweeping.
- To detect black-and-white thinking, ask: Does the statement claim to apply in all settings?
- This bias arises because we like shortcuts that can be learned and applied quickly. In the past, the speed of a black-and-white rule helped with life-or-death situations. Nowadays, such a rule can be applied at scale.
- Black-and-white thinking is incorrect if the world is:
  - *Moderate* : something is good only up to a point (water) or bad only after a point (carbs). However, we’re wired to think that more is always better or always worse, even if the data shows the opposite. Similarly, we view hitting targets as all or nothing,

when improvements and declines usually have gradual impacts.

    - *Granular* : something comes in many different forms, some of which are good (complex carbs), others bad (simple carbs). However, we extrapolate from individual trees to the entire forest. We engage in *categorical thinking* : we put items into buckets and base decisions on these buckets, not their individual contents.
    - *Marbled* : something contains both good and bad elements, such as semiconductor companies. However, we form an opinion based on a single positive or negative attribute. We classify a company as sustainable or unsustainable, ignoring its actual level of sustainability, or evidence as for or against, neglecting its actual strength.

Part I has discussed two biases that cause us to blunder when interpreting information. Confirmation bias applies when we have a prior view, and black-and-white thinking kicks in when we don't. These biases are mutually reinforcing – we're more likely to believe a claim if it's both appealing and extreme. We'll thus refer to them as the 'twin biases'.

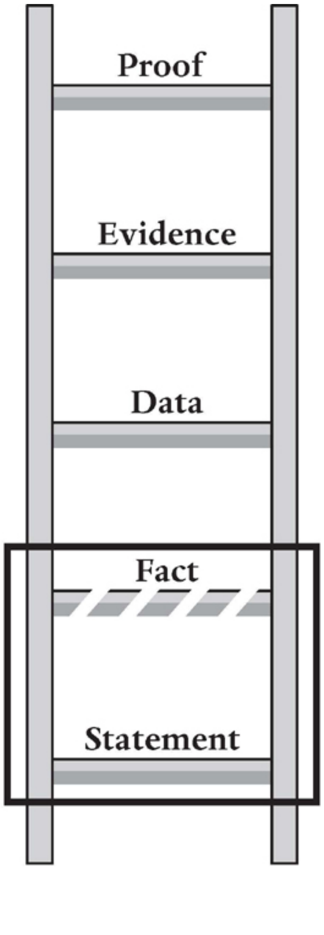
In Part II we'll explore the different problems that arise from the twin biases: we mistake statements for facts, facts for data, data for evidence, and evidence for proof. Highlighting these errors will help us be on our guard and know when to dig deeper. We’ll also introduce some simple questions that we can ask to ensure we don’t climb the Ladder of Misinference. The Appendix will later summarize them in a simple checklist.

PART II

The Problems

3

A Statement is Not Fact



It was a talk by a Cirque du Soleil performer that first introduced me to Malcolm Gladwell’s 10,000-hours rule. The acrobat, James, recounted how he’d developed the extreme skill required for his job. Friends thought he just happened to be born double-jointed or innately talented, but James stressed that he had no special gene; instead, his ability to perform one-armed handstands stemmed from hours of deliberate practice. James explained how Gladwell had proven that anyone – regardless of genetics or upbringing – can develop any skill, as long as they’re willing to spend 10,000 hours working on it.

It was a powerful message and one that I was eager to believe. From an early age, parents, teachers and well-meaning family friends tell kids ‘You can do anything you set your mind to’ and ‘Practice makes perfect.’ I’d repeatedly heard the same mantras, so I was primed to accept this rule. It makes you feel not only empowered, but also smug – you can proclaim that any success is down to your hard work rather than the lottery of genetics.

The rule also plays into black-and-white thinking. Beyond the general idea that more practice is always better, it suggested that any type of practice helps, even if the world is granular and some forms are better than others. I suppose at some level I’d always believed in a principle like this. During my time at MIT, I played tennis with Florian Ederer, a fellow Ph.D. student I’d known since our undergraduate days at Oxford. He kept bugging me to keep score and play actual matches, but I felt my Ph.D. was already stressful enough so I just wanted to knock the ball around. (Plus, I was a bit intimidated at keeping score against an opponent named F. Ederer.) I thought that it didn’t matter how we spent our time – as long as we were on the court, it would count as an hour of practice.

Yet I couldn’t just take James’s word for it. I needed to research the rule myself. I found a paper in a peer-reviewed medical journal which stated that ‘the 10,000-hour rule assert[s] that the key to achieving true expertise in any skill is simply a matter of practicing, albeit in the correct way, for at least 10,000 hours’.[1] According to a *Fortune* article, the rule explains ‘why some of us aren’t destined for success – but can still make it big’. And this article contained Gladwell’s own words: ‘The 10,000-hours rule says that if you look at any kind of cognitively complex field, from playing chess to being a neurosurgeon, we see this incredibly consistent pattern that you cannot be good at that unless you practice for 10,000 hours.’[2]

Unambiguous as it was, a magazine quote wasn’t enough for a self-respecting academic like me. I bought a copy of *Outliers*, the book in which Gladwell introduces the rule. In it, he stresses how ‘Ten thousand hours is the magic number of greatness,’ ‘researchers have settled on what they believe is the magic number for true expertise: ten thousand hours’ and ‘Practice isn’t the thing you do once you’re good. It’s the thing you do that makes you good.’ Gladwell also explains the science behind the rule, calling ‘Exhibit A . . . a study done in the early 1990s by the psychologist K. Anders Ericsson and two colleagues at Berlin’s elite Academy of Music.’

According to Gladwell, the researchers divided the academy’s violinists into three groups – the ‘best’ violinists with the potential to become world-class soloists, ‘good’ violinists, and those who’d end up as music teachers. By the age of twenty, the future teachers had practised for 4,000 hours and the good violinists for 8,000. What about the best? You guessed it – 10,000 hours. Beyond that study, Gladwell gives examples ranging from Sun Microsystems founder Bill Joy to Mozart – ‘even Mozart – the greatest musical prodigy of all time – couldn’t hit his stride until he had his ten thousand hours in’.

I had all I needed. Not only did I believe the rule myself, but I confidently taught it every year to my Wharton students from then on. Armed with this rule, I transformed myself from a mere finance professor into an aspiring motivational speaker and life coach. I stressed how my students could learn any skill they wanted to during their MBA – public speaking, negotiation, or cross-border differences in accounting for intangible assets – if they just believed in themselves and put in the hours. In response, I’d get knowing nods, so I was reassured that this was a truth universally acknowledged.

A few years later, after I’d moved to London Business School, I took up a part-time position at Gresham College – an unusual institution that doesn’t offer degrees but only free lectures to the public. One of my lectures was on the ‘growth mindset’: the idea that people aren’t born with a fixed set of skills but can develop them through hard graft. The 10,000-hours rule would take centre stage in that lecture, whereas I’d only mentioned it briefly to my MBA students. So I decided to read *Outliers* again, this time with a fine tooth-comb.

To my dismay, I found that it said something subtly, but critically, different from what I’d been teaching. The quotes from others had implied that 10,000 hours are *sufficient* for sucess – ‘the key to achieving true expertise in any skill is simply a matter of practicing’, so even if you had no talent, you could get to the top with toil and sweat. But Gladwell only claimed that they are *necessary* for success; you need both practice *and* talent. Rather than promising that you will be good if you practice, he warned that ‘you cannot be good . . . unless you practice’.

Yet Gladwell is far from blameless either. Stung by this misreading, I realized I needed to scrutinize the Ericsson paper also. True enough, the study was about violinists. Ericsson and his co-authors asked each student to keep a diary of how much time they devoted to deliberate practice. But in contrast to Gladwell’s claim that the best violinists practised more than the good ones, the researchers found *no difference* – each averaged 24.3 hours per week.

Disappointed but undeterred, I read on, biasedly searching for the evidence that would allow me to keep teaching the rule. The researchers asked each budding musician, now aged twenty-three, how much weekly practice they'd done each year since starting the violin, which for most students was aged five. Five! How accurate could their recollections possibly be, eighteen years later? I recalled my own experience as a five-year-old playing chess, which I probably took as seriously as the violinists back then. I was often asked how much I practised but could never give a number; it varied from week to week, depending on schoolwork and upcoming tournaments. Estimating just a ballpark was difficult, even in real time – let alone eighteen years afterwards. With a growing sense of apprehension, I had to admit to myself that the study relied on rough guesses, in contrast to Gladwell's definitive statements.

Most damning of all, there was no mention of 10,000 hours in any table or figure. The only number quoted for the 'best' violinists was this: 'by age 18 . . . the best young violinists had accumulated an average of 7,410 hr of practice'. The study did contain a graph of the hours of practice amassed by each group until the age of twenty. Eyeballing the graph, the best students averaged somewhere between 10,000 and 11,000 hours by that age. This number was indeed more than the 'good' group, which in turn exceeded the future teachers.

But why age twenty? Ericsson clarifies in his own book, *Peak*, in a chapter titled 'No, the ten-thousand-hour rule isn't really a rule', 'at eighteen or twenty, these students were nowhere near masters of the violin . . . Pianists who win international piano competitions tend to do so when they're around thirty years old, and thus they've probably put in about twenty thousand to twenty-five thousand hours of practice by then.' The graphs for all three groups increased with age, so even the future teachers would reach the magic 10,000 number – if Gladwell was right, they too would become world-class. Sure, it might take them a bit longer, but they'd get there in the end. Wasn't the whole power of the rule that it only stipulated a number, not a timescale – which is why practice is 'the thing you do that makes you good'?

This episode taught me an important lesson: *a statement is not fact*, because it may not be *accurate*. I'd heard about the rule from James, from a scientific paper, and from a *Fortune* interview, and latched on to a few sentences in *Outliers* which confirmed their version. But this wasn't enough: I needed to read Gladwell's book with a fresh mind, forgetting about what other people had told me. And even that wasn't enough: I needed to comb through the study Gladwell quoted and make sure it actually supported his claim. Note that I didn't need statistical smarts to do so – I just needed to be able to read English – but I never bothered to check the statement, because I wanted it to be true.

I was embarrassed that I'd taught the popular version of the rule to my students for many years, spreading misinformation. But was I being too hard on myself ? My talks never put any special emphasis on the 10,000-hours figure; instead, they were about the general idea that practice is important. Nor did I give the impression that ability didn’t matter; I was encouraging my students to work on public speaking, not become ballerinas. And are we being too harsh on Gladwell? Is it really so bad to highlight the merits of hard work? Does it really matter if a statement is not precisely fact?

Putt-ing it into practice

Dan McLaughlin was inspired. Aged thirty, his life up to that point had been one of drifting – he’d switched colleges, lived in multiple states and worked various jobs as a commercial photographer. Then he heard about the 10,000-hours rule and was galvanized by the idea that you can become great at something – anything – simply through honest effort. Golf seemed like the perfect sport to test the theory. It’s mentally and physically demanding, although not as forbidding as other sports that require you to start young, and he’d enjoy the challenge. It’s also objective, with a handicap system that would allow Dan to measure his progress.

Dan set his sights high, with a dream to make the PGA tour. He hadn’t even played a full round of the game, but the 10,000-hours rule was his ticket to glory. He saved up, quit his job and started to climb the mountain towards the magic number. Dan sheepishly began his training at the public golf course in Portland, Oregon, first taking baby steps. He stood just one foot away from the hole, learning how to make a simple putt. He practised for four hours a day, every day; as his skills slowly developed, Dan moved further back from the hole, a few feet at a time. For four and a half months, all he did was putt. Then he graduated to hitting the ball into the air; before long, his four hours of daily practice would end with a full round of eighteen holes.

After two years of non-stop effort, Dan’s handicap was 8.7; it became 6.2 after three years and dropped further to 3.3 at the end of four. For most of his fifth year, his handicap hovered around 3, putting him in the top 4% of the 26 million golfers in the US. That was far from world-class – there were 1 million Americans at least as good as him – but still pretty good.

Just after the 6,000-hour mark, disaster struck. Out on the course one day in the middle of hitting a drive, Dan felt a sudden pain in his lower back. He couldn’t take another shot and was barely able to walk off the course. A doctor told Dan he’d slipped a disc. Dan tried several times to resume after weeks of rest, but each time it only took a few holes for the pain to flare up again. Five months after his injury, he had to give up on his dream.

It’s impossible to attribute Dan’s injury to a single cause, but excessive training may have played a big role. The obsession with hitting 10,000 hours led Dan to over-practise in a sport that involves repetitive motions – particularly since the rule stresses how you should focus only on the skill in question, rather than cross-training through weights or yoga. And Dan’s losses were more than just the injury: he ‘cast everything else aside – career, money, even relationships’ to pursue his goal, according to the *Atlantic*. This included investing all the money he’d set aside for graduate studies.

Inspiring realistic, even ambitious hope is to be encouraged. But by insisting that anyone can be an expert, the popular form of the 10,000-hours rule leads people to fritter away their time and money chasing castles in the sky. Even Gladwell’s version dupes them into focusing on the quantity, not quality, of practice. Gladwell claims that *any* time rehearsing the activity counts. He argues that The Beatles became successful because they were invited to play in Hamburg, where sets lasted up to eight hours, unlike Liverpool which only had sixty-minute gigs – the subtitle to the chapter ‘The 10,000-hour rule’ is ‘In Hamburg, we had to play for eight hours’. However, Ericsson’s study focused on solo, deliberate practice guided by a coach and explicitly separated out performing into a different category.

A final effect of the rule may be to dishearten, rather than encourage. Ten thousand hours is an enormous amount of time – twenty hours a week for ten years. It delivers the black-and-white implication that practice is futile *unless* you can reach 10,000 hours. If true, none of my MBA students should ever bother developing finance, public speaking or negotiation skills, even if they had innate talent. If ‘you cannot be good . . . unless you practice for 10,000 hours’, why even try?

There’s no problem with presenting a hypothesis. It’s certainly fair to tell the tales of Bill Joy, Mozart and The Beatles and speculate that practice lay behind their success. But that’s how we should view them – as masterfully told stories and innovative conjectures. There’s a vast chasm between proposing a hypothesis and declaring a rule that applies to ‘any kind of cognitively complex field, from playing chess to being a neurosurgeon’, claiming that you’ve successfully tested it,[3] and calling ‘Exhibit A’ evidence that doesn’t actually support it.

Like many examples in this book, those making missteps up the ladder probably didn’t set out to deceive their readers. They’re human, so they succumbed to the twin biases – just as I did when I taught the 10,000-hours rule without reading the Ericsson study. It’s because these biases are so pernicious that we need to take extreme care when climbing the rungs.

Did I actually say that?

You might not be too shocked that some books play a little fast and loose with the evidence. After all, the author’s incentive is to write what sells, not what’s true; as the saying goes, ‘Never let the truth get in the way of a good story.’ Unfortunately, the problem of misquoting statements and presenting them as facts applies even to official government reports, which should be the bastion of accuracy.

The UK government’s Select Committee on Business, which we came across in the Introduction, launched a separate investigation into executive pay in 2018. Their concern was that CEOs were being paid massive sums despite poor performance. A big chunk of my work is on how pay should be reformed, so I wanted to contribute and sent in evidence. I was curious to read the final report when it came out the following year.

After reading all the submissions, the Committee’s conclusion hadn’t changed from their initial hunch that CEOs are overpaid. Key to this view was the belief that a single CEO doesn’t matter as there are thousands of other employees in a firm – the report stated: ‘the evidence is at best ambiguous on the impact of individual CEOs on company performance.[110]’ However, a large body of scientific research finds the reverse. The wider workforce certainly matters, but so do CEOs. It’s not either–or, just as both football managers and players make a big difference to a team’s success.

The report referred to footnote 110, so I glanced at the bottom of the page to see who’d had the temerity to lie in a government inquiry. My jaw dropped when it said ‘110 Professor Alex Edmans’. Flustered, I pored over my initial submission, worried I’d made a fatal typo – but my evidence clearly stated the opposite: that CEOs have a significant impact on firms.[*] The Committee read what it wanted to read.

I'm willing to believe that the Committee didn't deliberately misrepresent my submission. But as we saw in Chapter 2, if you have a strong prior belief, confirmation bias leads you to interpret *any* evidence as being consistent with it, even if it's ambiguous or contradictory. This misinformation then spreads. Many readers will have noticed there was a footnote, reassured themselves that there must be evidence behind it, and believed it without examining who the evidence came from. Others may have checked the footnote, seen it was by 'Professor Alex Edmans', kindly thought that I must be semi-reliable since I work in executive pay, and similarly moved on. That's also insufficient – a reader needs to make sure that the reference actually supports the statement. Just because there's a footnote at the end of a sentence, it doesn't mean the sentence is true.

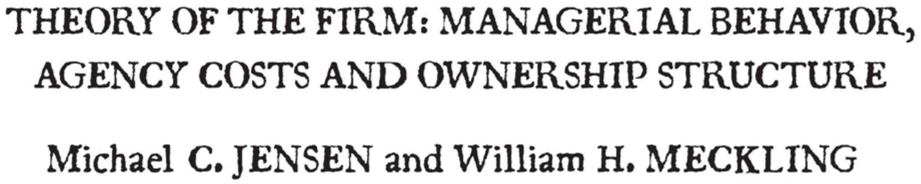
Garfunkel & Simon's greatest hit

Gladwell's quoting of Ericsson shows that, even if an article is accurate, others can garble it to suit their own objectives. But it takes effort to notice the fudging – you have to dig into the weeds of a paper to find out what it actually measured, tested and concluded. My submission to the UK government inquiry was eleven pages of 4,500 words, and someone who finds executive pay less captivating than I do may not have the stomach to wade through the swamp.

Fortunately, you might not need to look past the first page to spot a misportrayal. All academic studies begin with an 'Abstract', a 100–150-word plain-English summary of the main results. For the paper on pay gaps mentioned in the Introduction, the abstract was loud and clear: 'firm value and operating performance both increase with relative pay'. And sometimes, you don't need to read even that.

An opinion piece in *Forbes* by businessman Steve Denning argues that 'In one of the most-cited, but least-read, business articles of all time, finance professors William Meckling and Michael Jensen offered a quantitative economic rationale for maximizing shareholder value[†] . . . Meckling and Jensen proposed a licence for enterprises to pursue unbridled self-interest across an entire society.' Similarly, Simon Sinek's book *Leaders Eat Last* claims that 'William Meckling . . . and Michael Jensen' came up with the 'answer everyone was looking for', 'a simple metric for measuring corporate performance' – shareholder value.

Denning was right about one thing – the article being not widely read (there's no evidence for 'one of the . . . least-read', an assertion that plays into black-and-white thinking). He himself may not have read it, because it starts with this:[‡]



THEORY OF THE FIRM: MANAGERIAL BEHAVIOR, AGENCY COSTS AND OWNERSHIP STRUCTURE

Michael C. JENSEN and William H. MECKLING

The article is by Jensen and Meckling, not Meckling and Jensen. Am I being nit-picky about getting the authors' names the wrong way round? In fact, it's a valuable red flag. It gives away how Denning and Sinek never bothered to open the paper; instead, they just pulled up Wikipedia. The Wiki entry for 'shareholder value' at the time read: 'finance professors William Meckling and Michael C. Jensen . . . provided a quantitative economic rationale for maximizing shareholder value' – almost exactly what Denning wrote. It seems Denning wanted to attack shareholder value, looked it up on Wikipedia, and copy-and-pasted the description of a study without reading it.[§4] If he wished to indict the research for society's ills, he needed to have scrutinized it carefully, yet he hadn't even glanced at page one. It's like claiming you're a massive fan of Garfunkel & Simon, or you're so attractive that you model for Fitch&Abercrombie.

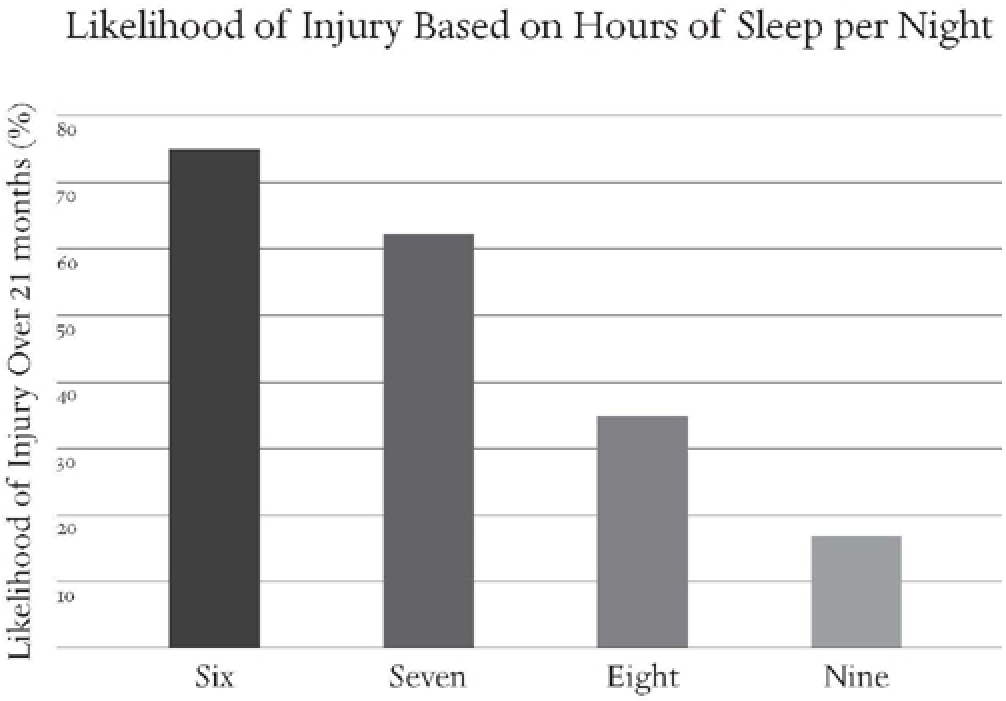
In fact, Jensen and Meckling didn't come close to recommending 'unbridled self-interest'. Their very first graph has two axes, one for shareholder value and the other for 'non-pecuniary benefits', and shows that a company should be in the middle. And what do 'non-pecuniary benefits' comprise? Things like 'charitable contributions, personal relations ("love", "respect", etc.) with employees' – hardly uninhibited greed. Yet hordes of readers may have accepted Denning's and Sinek's black-and-white statements because they confirmed their view of capitalism.

Here, getting the researchers' names the wrong way round was a glaring indicator that the authors hadn't read the study they were citing. Other articles misquote the title or miss out an author. Just glancing at the header of a study, without even getting to the abstract, is sometimes enough to detect a distortion.

Choosing your words (and data) carefully

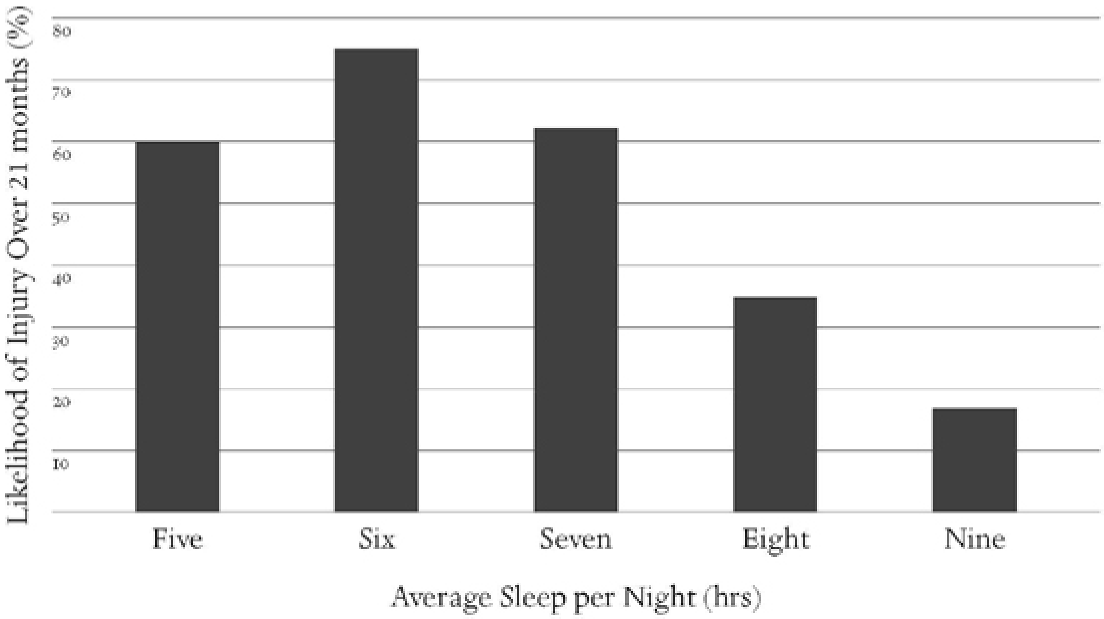
There's a different shortcut if a statement is a direct quote – you can simply search for it without having to trudge through the whole report. Thousands of articles claim that former General Electric CEO Jack Welch declared that 'Shareholder value is the dumbest idea in the world.' Google quickly tells you it's from a *Financial Times* interview, and a Ctrl-F on that interview reveals the full quote as 'On the face of it, shareholder value is the dumbest idea in the world. Shareholder value is a result, not a strategy' – which has a quite different meaning. So while the reference wasn't technically false, it's still a lie because it's selective and out of context. Those citing it chopped off the head and the tail to leave a black-and-white body.

Selective quoting can also occur with data. Matthew Walker’s bestselling 2017 book, *Why We Sleep*, presents a bar chart showing how more sleep is associated with fewer injuries in teenagers.



Average Sleep per Night (hrs)

Likelihood of Injury Based on Hours of Sleep per Night [#]



Walker cites this graph as being from a paper entitled ‘Chronic lack of sleep is associated with increased sports injuries in adolescent athletes’.[5] This graph, and Walker’s whole book, hits our twin biases right between the eyes, as we’d all like an excuse to stay in bed longer and to think that over-eager beavers will get their comeuppance. Since this is a specific chart, we can easily check it without having to read the full article. The four-and-a-half-page study has a single graph which looks like the one at the bottom of the page opposite:[**]

Walker removed the bar showing that five hours’ sleep leads to fewer injuries than six or seven because it doesn’t fit his message. That’s like the police hiding evidence that might exonerate their suspect.

In all these cases, it's easy to get to the truth. You don't need any knowledge of statistics; all you need is to look up the original source. If it's research, confirm what the authors themselves concluded; if it's a quote, check the person said it along with the context; if it's a graph, inspect the actual graph. Of course, you'd go crazy if you went to the horse's mouth each time you saw a reference. However, if it's a particularly important claim and it plays into the twin biases, it's especially valuable to be sure.

Sometimes, you may not even need to go to the source – for an influential book or article, someone else may have played detective. Googling 'Why we sleep criticism' or '10,000 hours criticism' (without quotes) will uncover myriad problems with both books, over and above the ones we've discussed. Yet if confirmation bias is at play, we'll accept a thesis at face value and not bother to look for any flaws.

All these examples illustrate the first step in verifying whether a statement is fact: *check if the authors' conclusions match the statement*. Doing so helps us guard against a third party misquoting a study, evidence submission or interview to suit their own purposes. But the problems don't just lie with third parties – the authors themselves may be to blame. We'll now turn to an example, and one that hits uncomfortably close to home.

The evidence that wasn't there

20 July 2021 started off a proud day for me. London Business School, where I'm a professor, released a report claiming that boardroom diversity improves firm performance. This wasn't just any report but one commissioned by the Financial Reporting Council, the UK regulator, so it would likely shape practice. As a member of an ethnic minority and a supporter of diversity of all forms, I was excited to read the press release entitled 'Diverse boards lead to better corporate culture and performance'.

This finding seemed a no-brainer: by having a range of viewpoints, we can overcome individual biases – perhaps even the twin biases – and make better decisions. While the idea sounds obvious, it's important to test it with data, and this is what LBS had done.

My experience with the 10,000-hours rule had taught me not to take a quote for granted, so I began the marathon of reading the 132-page study. The Executive Summary backed up the press release, highlighting that 'Higher levels of gender diversity of FTSE 350 boards positively correlate with better future financial performance (as measured by EBITDA margin)', a measure of profitability.[††] So far, so good. However, that's only a statement about the results; it's not the actual results. For those, I needed to delve into Tables C.7, C.8 and C.9 in Appendix C, which link gender diversity to the EBITDA margin. These tables contained ninety different tests – and every single one of those ninety tests found no relationship.[6] The authors announced a result that just wasn't there.

This misrepresentation is not only disingenuous but unnecessary. Most companies pursue diversity because it’s the right thing to do, not to make money. And even if there’s no link between diversity and performance, that would still support diversity initiatives as it suggests that you can increase diversity for free, without having to sacrifice profitability. Yet many newspapers, companies and professional bodies accepted the claims unquestioningly, writing headlines such as ‘Boardroom diversity improves financial performance’.[7]

This illustrates a second reason why a statement is not fact: authors may misrepresent their own findings because they’ll be more influential if they have an appealing punchline. It’s not enough to verify that a third party’s statement is backed up by the authors’ conclusions. We need to take a second step: to *check if the authors’ results match their conclusions*.

And we need to go further still. The third step is to *check if the authors’ data matches their conclusions*. A study may misportray not just whether the data shows a link but what the data is actually capturing. It might measure something quite different, miss a big piece of the picture, be self-reported or be circular.

As an example of where *the data measures something different*, after George Floyd’s murder in May 2020 there were numerous Black Lives Matter protests to fight for racial justice. Some worried that mass protests might increase the spread of COVID-19, but a study claimed that ‘cities which had protests saw an *increase* in social distancing’.[8] It was widely covered by the media, because their readers wanted this result to be true. Yet the researchers didn’t measure social distancing but the amount of time spent at home (captured by mobile-phone location). Social distancing depends on your distance from other people, not your home. You might leave the house for only two hours, but if it’s to participate in a mass rally, that’s not social distancing.

This misrepresentation is again unnecessary. Many people believe that protesting against centuries of inequality is justified even if it reduces social distancing, just as you should still take a medicine despite its side-effects. But black-and-white thinking means we want to label something as unambiguously good or bad. As a result, a study is less likely to go viral if it suggests there might be trade-offs, no matter how minor.

A second problem is when *the data misses a big piece of the picture*. In 2022, a study argued that banning the advertising of unhealthy food on London public transport would stop 100,000 people from becoming obese and save the National Health Service £218 million.[9] The Mayor of London praised the research, but that’s not surprising, since he imposed the ban. I’m a health and fitness maniac, and as a behavioural economist I believe that nudges like advertising make a difference, so I was ready to lap this one up.

Yet looking beyond the headline to page three of the report, you can see that they measured 'food and drink items purchased and brought into the home' – but a huge chunk of unhealthy food is consumed outside the house.[10] Their numbers completely missed burgers at McDonald's, beer down the pub, and more burgers and more beer at the football.

A third source of inaccuracy is if *the data is self-reported*. In my first book, *Grow the Pie*, I wrote positively about Share-Action, a charity which harnesses shareholder power to make companies more responsible. In July 2022, they led a campaign to force the supermarket Sainsbury's to pay the Living Wage.[‡‡] The moral justification for paying living wages is already compelling. Yet ShareAction added a different argument – it claimed there's widespread evidence that living wages would improve Sainsbury's profits.[11] Rather than taking pie away from shareholders and giving it to workers, increasing salaries would grow the pie – employees would be more productive and likely to stay, ultimately benefiting shareholders.[§§]

Given this was the argument in my book (even though ShareAction didn't use my specific metaphor), I was eager to embrace it, but first looked up one of the studies they mentioned.[12] It never actually measured profitability but simply asked companies that became Living Wage employers whether they *thought* it had boosted their bottom line. Most claimed it had. But as Mandy Rice-Davies, who gave evidence in the 1963 trial that discredited the government of UK Prime Minister Harold Macmillan, is commonly paraphrased: ‘They would say that, wouldn’t they?’ – often abbreviated as MRDA (‘Mandy Rice-Davies applies’).[‖‖] If you’ve taken a major business decision, you’re bound to argue it was the right one, due to confirmation bias. Whenever data is self-reported, people might say whatever they want to be true.

The final problem is when *the data is circular*. Almost every study investigates how an *input* affects an *output* – whether the death penalty deters crime, water improves marathon performance, or junk food bans curb obesity.[##] Sometimes the input and output measure pretty much the same thing, so any relationship is automatic. A 2020 McKinsey report on company responses to COVID-19 found that resilient companies outperformed – their sales and profits held steady, while nonresilient ones saw them drop off a cliff.[13]

How did McKinsey define the resilient companies? Not by CEOs rolling up their sleeves and mucking in, nor by a gritty corporate culture, but as the ones whose share price performance was in the top 20%. Sales and profits are two of the most important drivers of the share price, so the top 20% will have almost automatically had strong sales and profits. That’s as astonishing as finding that ‘football teams that win more games score more goals’. Yet by labelling good stock price performance ‘resilience’, a quality people admire, the study was able to make the alluring claim that resilience pays off.

The study that never was

Is there anything worse than bad data? There is – having no data at all. In 2019, two influential organizations jointly launched a study[14] with the press release ‘CEO remuneration packages actively discourage innovation in UK’s top companies’. Yet there wasn’t a single test to support this result. All they did was gather executive pay data, show it has certain features such as bonuses, and *assume* that bonuses discourage innovation – there was data on the input (pay) but not the output (innovation). As the Queen of Hearts scoffed in *Alice’s Adventures in Wonderland*, ‘Sentence first, verdict afterwards.’ They’d decided that executive pay needs to be punished, without giving evidence of any guilt.

Surely this section ends here – there can’t be anything crazier than having a paper with no data? There can – having no paper. In March 2022, Reuters published an article headlined ‘Boardrooms with more women deliver more on climate’,[15] referencing a study by a leading investor, Arabesque. I was intrigued, given my interest in both diversity and climate, but there was no such study. The article didn’t link to any paper, and there was nothing on Arabesque’s website; a colleague emailed Arabesque, but they never replied. The Reuters article was all over my LinkedIn; when I asked those sharing it if they could send me the study, they admitted they hadn’t actually seen it but had taken the journalists’ word for it. Yet if an article described an unwanted result, readers wouldn’t immediately accept it but demand to see the analysis so they could pick it apart.

Likewise, a newspaper will sometimes write about ‘a study exclusively shared with the *Daily News*’, giving the impression it’s letting you in on some big secret. Instead of being highly excited about the article, we should be highly sceptical – if the authors won’t publicly release the research, it probably has gaping holes. It’s like a musician claiming to have recorded the next ‘Gangnam Style’ but not letting anyone hear it. If his song really were so catchy, he’d let people judge for themselves.

As with the misquotes and the artfully edited graphs, the solutions aren’t difficult. You can check a study exists with a quick web search. If it does, see whether it’s conducted the analysis it claims, and if so what data was used and whether the results actually found a link. It’s impractical to scrutinize every paper in this way, but we can be selective. With some studies, there’s little ambiguity. For research linking a football club’s wage bill to its league position, or a person’s hours of sleep to their body mass index, we can be pretty sure how the authors measured their inputs and outputs. For something trickier like social distancing, consumption of unhealthy food or resilience, we do need to check how these concepts were captured – particularly if confirmation bias means there’s a preferred result.

Out of the blue

After decades of attempts to enact healthcare reform in the US, the Democratic Party made another push in July 2009. Seven Democratic Representatives proposed HR 3200, the America’s Affordable Health Choices Act. Among other things, it would require all citizens above poverty level to buy private insurance, with government subsidies for low-income households, thereby ensuring access to healthcare for all Americans. The bill later morphed into the Affordable Care Act, commonly known as Obamacare, which was signed into law the following March.

Few could have anticipated the political storm that Section 1233 of HR 3200, which took up a mere ten of its 1,017 pages, would ignite. This section suggested reimbursing doctors for consultations to Medicare patients about end-of-life-care options and ‘living wills’ in which people set out how they’d like to be treated if they later became incapacitated. The proposal was innocuous as the sessions would be voluntary; all it meant was that patients would be entitled to a free consultation every five years and doctors would be paid for providing it.

But opponents claimed it would introduce a ‘death panel’ of bureaucrats who’d decide whether Americans, particularly the elderly and children with disabilities, were worthy of healthcare or instead should be encouraged to die. Former Vice-Presidential candidate Sarah Palin wrote on Facebook that ‘The America I know and love is not one in which my parents or my baby with Down Syndrome will have to stand in front of Obama’s “death panel” so his bureaucrats can decide, based on a subjective judgment of their “level of productivity in society,” whether they are worthy of health care. Such a system is downright evil.’

This fear quickly spread, and 30% of those who heard the claim believed it, with 20% undecided.[16] Yet there was nothing in Section 1233 that suggested anything remotely close to a ‘death panel’. If there had been, the best way to call out the ‘evil’ would have been to quote the offending clause, but no critic ever did so. This absence should have rung alarm bells, but instead the myth was naïvely accepted. Not only did it cause many politicians and voters to oppose the Act, despite its many benefits, but it also diverted public debate from legitimate concerns such as the cost of subsidized insurance. The political fact-checking website PolitiFact named death panels the ‘Lie of the Year’ for 2009.

So far, we’ve seen how people misrepresent the conclusions, the data analysed and whether there was any data or even a study. The death panel episode is a final example of why a statement is not fact: some statements can be completely made up. They’re simply fabricated out of the blue, not a reference in sight – but if they’re extreme, people often believe them without asking for proof. As the saying goes, ‘You couldn’t make it up,’ so they think it must be true.

Some statements mention no source at all, not even a vague reference like the one to Section 1233. Senator Bernie Sanders claimed: ‘Wall Street CEOs who helped destroy the economy, they don't get police records. They get raises in their salaries.' He didn't give any evidence, and people didn't ask – it seems so outrageous to be rewarded for crushing your company that it must be correct. In fact, the CEOs of both Bear Stearns and Lehman each lost nearly $1 billion when their companies went under.

When statements don't have sources, the checks we've previously mentioned are harder to conduct.[***] What do we do in these situations? There's little we can do actively – we can't check the sources if none are given. However, we can respond passively by putting much less weight on a claim if no evidence is provided, particularly if the claim is extreme. The statements by Palin and Sanders were conspicuous in the lack of their evidence despite the strength of their assertions.

Statements that can never be facts

Most of this chapter has discussed simple yes/no statements that can be easily proven or disproven. Section 1233 either proposes death panels or it doesn't; Lehman and Bear Stearns bosses were given pay rises or they weren't; and living wages either improve company performance or they don't. A specific claim can be backed up by a sentence in Section 1233, a pay disclosure or a study.

But many statements we encounter day to day are about general ideas which can't be classified neatly into true or false; instead, they lie on a spectrum from more to less reliable. Examples include 'Religious extremism is the single biggest threat to the world,' 'Our country's education system is why we're behind our neighbours,' and 'As a driver of company success, culture eats strategy for breakfast.' There's no scientific study that can either prove or disprove these assertions beyond doubt. Yet this doesn't mean that we should ignore them. Our understanding of the world would be much poorer if we never looked beyond what we can certify with certainty – as Chapter 8 will explain, you almost never have 100% proof. So how much faith should we put in statements like these? Three principles we've already discussed continue to apply here.

The first principle is to *check if the statement plays into the twin biases*. Tariq Fancy, BlackRock's former sustainable investing chief, wrote an op-ed in March 2021 claiming that 'The financial services industry is duping the American public with its pro-environment, sustainable investing practices.' According to Fancy, sustainable investing does nothing to help the planet – it's like prescribing wheatgrass to a cancer patient – but allows companies to launch 'sustainable' funds and charge fat fees under the guise of doing good. The article went viral, and Fancy became a minor celebrity. He was praised as the paragon whistleblower who'd exposed the fraud of sustainable investing, granted op-ed columns in leading newspapers and invited to give prestigious conference keynotes – ironically charging fat fees himself.

All this fame would be deserved if Fancy had indeed exposed a scam. Was he right? You can't prove whether sustainable investing is good or bad for society – it has both positive and negative consequences, and reasonable people can disagree on whether the home runs outweigh the strikeouts. So let's apply our first principle. Fancy's claims fed on confirmation bias, because many people will eagerly believe that sharp-suited fund managers are crooks, scamming pensioners who want to do good with their savings. Fancy also exploited black-and-white thinking, alleging that the entire sustainability industry was a ruse, even though he only had experience at one company, BlackRock, and for less than two years. This doesn't mean his claims were definitely false – sometimes, what we suspect to be true is indeed true – but it does imply that the popularity of his views shouldn't be taken as a sign of accuracy.

The second principle is to *examine the evidence for the statement*. General opinions, like Fancy's, can't be confirmed or contradicted by a single study – but you can still provide suggestive evidence for them. Even if evidence can't turn something black, it can make it a darker shade of grey or a whiter shade of pale.

Fancy's article was short, so it didn't have space for stacks of stats. He followed it up with a forty-page essay on

Medium,[17] but it still contained no evidence, only anecdotes. I wrote a reply on Medium,[18] where I acknowledged that some of Fancy's points were valid but highlighted others that the data contradicted. The *Wall Street Journal* then invited us to a debate on 'Does sustainable investing really help the environment?'[19] It took the form of an email exchange, which the *Journal* then wrote up. The exchange was cordial; there was some common ground, and on other points we agreed to disagree. However, Fancy repeated his claim that 'there is now evidence emerging that [sustainable investing] may be a giant societal placebo that lowers the likelihood of us following expert recommendations to address the climate crisis'. When I asked to see this evidence, so I could either concede his point or respond to it, he couldn't provide any – yet he still maintained his claim to evidence in the final article.

The bar should be applied to both sides of any debate. Sustainability defenders were equally exaggerated, with a Global Head of Sustainability Research labelling sceptics as spouting 'just complete BS', and an Oxford professor calling them 'Taliban' and 'Flat Earthers'.[20] They were heralded as heroes by the pro-sustainability crowd, but because of the intensity of their indignation rather than the robustness of their rebuttal.

What if it's not appropriate to provide evidence? Let's assume Jack Welch's statement 'Shareholder value is the dumbest idea in the world' was taken in context and not preceded by ‘on the face of it’. That’s a general opinion for which there’s no proof. Nor could Welch have pointed to evidence, because he made that statement in an interview, not a forty-page essay. The key is to remember that such claims are merely opinion. ‘In Jack Welch’s subjective opinion, shareholder value is the dumbest idea in the world’ is a more accurate portrayal than ‘As Jack Welch stated, shareholder value is the dumbest idea in the world.’ Welch’s subjective opinion may still be valuable, since he’s a highly experienced ex-CEO, but this ensures we’re mindful it’s not fact.

Some statements involve superlatives, such as the ‘*dumbest* idea in the world’. In these cases, we can ask whether we can come up with a clear counterexample – the equivalent of ‘checking the facts’ for a statement that can be verified. For Welch’s quote, even the biggest critic of shareholder value could come up with several dumber ideas. Related to superlatives are universal statements. Venture capitalist Angela Strange quipped that ‘Every company will be a fintech company,’ which was catnip to the fintech crowd. But restaurants, theme parks and hardware shops are unlikely to ever become fintech firms. On the other hand, Sir David Attenborough’s warning that climate change is ‘our greatest threat’ is a superlative, but not clearly incorrect. Even though we can think of other threats, such as pandemics and nuclear war, they’re not obviously greater than climate change.

This counterexample check is useful because superlatives prey on black-and-white thinking. We recognize that Strange didn't literally mean that every single enterprise will become a fintech one but was just conveying the growth of fintech. Yet the fact that she chose to do so with an extreme statement may signal that the actual underpinning is weak. By claiming that 'shareholder value is the dumbest idea in the world' rather than 'shareholder value isn't always the best goal', Welch was able to attract huge attention despite providing no evidence.

The third principle is to *explore alternative explanations.* This applies especially to statements about cause and effect. Consider the claim by political scientist John Mearsheimer that NATO's eastward expansion provoked Putin to invade Ukraine.[21] There's no dataset that will prove or disprove this allegation. However, we can immediately check our biases and be mindful of rival theories that go against our knee-jerk reaction. If we're a NATO sceptic, we're eager to accept the claim, but perhaps other factors were behind Putin's decision. If we're pro-Western, we can ask ourselves whether NATO might have played a small role, even if it was far from the main contributor.

This principle is particularly useful for claims concerning ourselves, where our biases are the strongest. We're tempted to think, 'The boss promoted Andrea over me because they were at university together' when it's in fact due to our underperformance – but burying our head in the sand allows us to avoid facing up to reality.

In a nutshell

- *A statement is not fact* because it may not be *accurate*. People often use studies and quotes to support their point, but:
    - They may misrepresent what the study finds. A footnote at the end of a statement doesn't mean that the footnote actually supports the statement.
    - A quote may be out of context, such as citing only part of a sentence or cutting out a bar from a chart.
- Even if the study's conclusions match the person's statement, they still may not support it:
    - The conclusions misportray the results (for example, claiming a link when none was found).
    - The results match the conclusions but the data doesn't:
        - It measures something different (capturing social distancing by the amount of time spent at home).
        - It misses a big piece of the picture (ignoring junk food consumed out of the house).
        - It's self-reported (asking if companies thought they

        were successful).

        - The input is very similar to the output (measuring resilience by share price performance and claiming that resilience boosts financial performance).

    - There is no data; the authors assumed their results.
    - There is no study; the authors simply issued a press release.

- To check a statement, we can ask the following questions:
    - Is evidence quoted?
    - If yes, does it actually exist?
    - If yes, do the authors' conclusions match the statement?
    - If yes, does the authors' analysis (results and data) match their conclusions?
- Some statements are made without any evidence at all. Instead, they are given in an extreme way, to mask the lack of evidence by suggesting the point is so obvious that none is needed.
- Other statements can neither be proven nor disproven.

  In such cases, we should ask:

    - Does it play into the twin biases?
    - Is evidence given? If the statement contains a

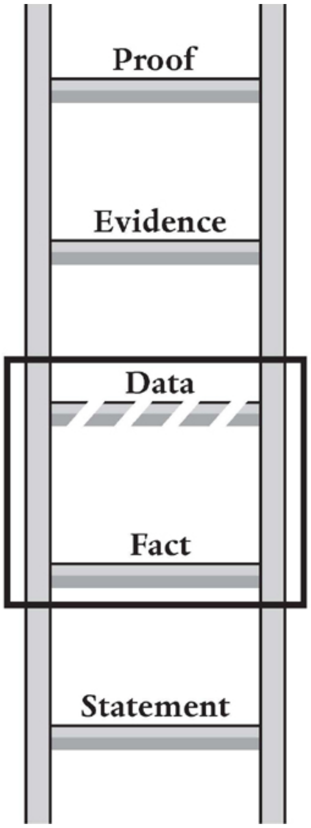
superlative, can we come up with a clear counterexample?

  - Are there alternative explanations?

But checking the facts isn’t enough. Even if a statement, story or statistic is accurate, it may still be misleading. The next few chapters will explain why.

4

A Fact is Not Data



Steve Jobs was born in February 1955 to Joanne Schieble and Abdulfattah Jandali, two students at the University of

Wisconsin. Joanne was from a strict Catholic family that refused to let her marry a Muslim man. To spare the family the scandal of a child born out of wedlock, Joanne's parents pressured her to flee to San Francisco to have the baby. Fearing she wouldn't be able to give her son the best start in life, Joanne gave him up for adoption.

Steve was adopted by Paul Jobs, a car mechanic, and his accountant wife, Clara. In 1961, the family moved to Mountain View, California, an area that would later become known as Silicon Valley. These humble beginnings planted the seeds for Steve's later success. His childhood home in Mountain View had been developed by Joseph Eichler, who'd used his formula of minimalist, elegant design – open-plan floors and glass walls – to successfully build over 11,000 homes in California. The young Steve grew up immersed in a modern style that could be mass-produced at scale.

His physical surroundings were complemented by a human touch. Paul applied his professional skills to DIY around the house, involving his son so that he could pass on his craftsmanship – a craftsmanship obsessed with appearance and perfection. As Steve later recounted: 'He loved doing things right. He even cared about the look of parts you couldn't see . . . For you to sleep well at night, the aesthetic, the quality, has to be carried all the way through.'[1] Paul refused to use cheap wood for the back of cabinets and made sure that the reverse of a fence was as expertly constructed as the front.

Steve’s childhood imprinted upon him the importance of design, which later became Apple’s hallmark. Thousands of companies aspire to become leaders in electronics, but they focus on functionality – what a product can do and how reliable it is. To Steve, that’s like building a cabinet to bear the greatest possible weight, but ignoring whether its aesthetic would turn your house into a home.

Instead, Steve’s passion was the *how* – how Apple’s products were made. This involved two elements: quality and simplicity. Apple goods were crafted with extreme care throughout, including in details invisible to the user. Apple’s engineers focused on getting the Apple Mac circuit board to work perfectly, but Steve agonized over what it looked like: ‘I want it to be as beautiful as possible, even if it’s inside the box. A great carpenter isn’t going to use lousy wood for the back of a cabinet, even though nobody’s going to see it.’ And he was fanatical about removing complicated features, even if they seemed indispensable. Any self-respecting BlackBerry user knew how essential the keyboard was. Yet Apple took that away with the iPhone, replacing it with a giant screen.

The *how* applied not only to how Apple approached its products, but also to how Steve approached life. His feeling of being abandoned by his birth parents drove him to prove himself, launching Apple aged just twenty-one. This determination, added to his penchant for design, meant that Steve not only valued craftsmanship but demanded perfection. And the *how* extended to how Steve approached setbacks. When the Apple board fired him in 1985, despite the company’s success, Steve was unfazed. He picked himself up and co-founded the animation studio Pixar. It was Pixar, not Apple, that made him his first billion. In 1997, Apple was floundering and Steve was reinstated as CEO. He then launched the iMac, iPod, iPhone and iPad – all undisputed successes that led to Apple becoming the first $1 trillion company in history, in August 2018. (It crossed $3 trillion in January 2022.)

While the *how* was the defining feature of Apple’s success, it was made possible by the *what*. Experts had argued for centuries that experimentation was the key to innovation. Most innovations fail, so you need to try lots of things to find one that works. Steve disagreed. When he returned to Apple in 1997, it was producing a dozen different versions of the Macintosh. To the shock of his team, he cancelled all but four. Steve drew a simple quadrant that defined the future of Apple: ‘Consumer’ and ‘Pro’ across two columns and ‘Desktop’ and ‘Portable’ down two rows.

This *what* was key to turning Apple around. In Steve’s words: ‘Deciding what not to do is as important as deciding what to do. That’s true for companies, and it’s true for products.’ It was the *what* that paved the way for the *how*. Concentrating on four products freed up Apple to pursue perfection in design.

Apple’s success was nothing to do with the *what* or the *how*. It was thanks to its *why*.

The Golden Circle Model stresses that *what* a company does is only its most superficial level. Deeper is *how* it manufactures its products or offers its services. But the centre of the Golden Circle is the centre of any company’s success – its *why*.

People don’t buy *what* you do or even *how* you do it. They buy *why* you do it. Customers queue up the night before the launch of a new iPhone, not because of its snazzy features or slick design but because they connect with why Apple makes iPhones. Apple’s *why* is ‘Everything we do, we believe in challenging the status quo. We believe in thinking differently.’ This inspires the hearts and minds of customers who are tired of being average, bored of just having what everyone else has and emboldened by the potential to be unique.

This Golden Circle isn’t just something made up by an ex-advertising salesman. It’s scientifically grounded in biology. The human brain is divided into three components which correspond perfectly to the three layers of the Golden Circle. The most superficial level is the neocortex, and it reflects the *what* – it’s based on rational, analytical thinking. Crucially, only *Homo sapiens* has a neocortex. Since humankind has only recently developed the neocortex, it plays little role in our behaviour. Instead, our thoughts, our words, our deeds – including those all-important buying decisions – are driven by the centre of the brain. That centre was shared by our ancestors millions of years ago. It’s primal.

And that centre is the limbic brain. It corresponds to the *how* and the *why* and is driven by emotion. Since it’s at the core, it’s responsible for all human behaviour, all decision-making – and, importantly, across all people. Some products succeed in the US but sit on the shelves overseas. Songs, books and films that are hits in Europe flop in North America. Apple became a worldwide success because it connected with the limbic brain, which characterizes *all* humans – young and old, male and female, black and white. It started with *why*.

You’ve just read two of the most famous explanations for Apple’s success. The first is from the multimillion-selling biography of Steve Jobs by Walter Isaacson. The second is from Simon Sinek’s ‘How great leaders inspire action’, the thirdmost-watched TED talk of all time, with over 60 million views and the basis for his book *Start with Why*. Yet these two accounts are completely different from one another.

But it’s their similarities that help us understand how two inconsistent explanations became lodged in Apple lore.[*] Both prey on black-and-white thinking. Adoption may have negative as well as positive effects, and launching a company based on a *why* rather than a business plan might seem like building a house on sand. However, if we’re predisposed to think something is unequivocally good or bad, we accept the explanation and ignore moderation or marbling.

Both accounts also play into confirmation bias. We like underdogs, so we root for adopted kids. We want to believe that a *why* leads to success because that's empowering. Perhaps Apple's fortunes instead stemmed from a eureka moment or a stellar network of contacts – yet many companies have neither, so any book claiming they're the key is unlikely to become a hit. But anyone can come up with a *why* if they think hard enough, hold long enough brainstorming sessions or hire expensive enough consultants. Believing Sinek puts you in the driving seat.

A third similarity is that both authors tell a compelling narrative. They make it seem like Apple's success was logical – predestined, even – given the company's *why* or its CEO's childhood. To do this, they reverse-engineer a story for Apple's success and make it vivid, exciting and appealing. If they only need to explain a single example, authors are free to concoct whatever account they want. They can then highlight some choice factoids to support it and ignore anything that doesn't fit. Neither considers the possibility of alternative explanations, which is why two highly popular accounts don't even acknowledge each other.

We've seen how the twin biases can lead us astray, but why is reverse engineering a problem? Is it really so bad to come up with an explanation after the event? Jobs himself didn't think so. He explained in a famous 2005 graduation speech at Stanford University that 'You can't connect the dots looking forward; you can only connect them looking backwards.' Shouldn't we use the benefit of hindsight, when all the facts are on the table – to see all the dots before we try to join them? And why do we complain about brushing aside what doesn't match the narrative? Isn't the skill of an author, a journalist or even an academic researcher to see clarity in chaos: to tune out the noise and focus on the signal?

To understand the flaws in this popular approach, let's dip into one of the most influential streams of finance research at the turn of the millennium.

Seeing the full picture

Reyes the Entrepreneur makes it sound so simple. His YouTube videos, 'How I made $4,000 PROFIT investing in STOCKS', 'How much investing in stocks made me in 24 hours' (answer: 3.311% in a single day), and 'How much A DAY TRADER makes in ONE DAY', explain with enthusiasm and passion – and capital letters – how easy it is to make money on the stock market. With such grand promises, it's not surprising his videos have been viewed over 40 million times. And it's not just Reyes. If you put 'how to make money in stocks' into YouTube, Google or Amazon, you'll find no shortage of videos, articles and books suggesting that playing the market is as easy as playing the triangle. If you're in it, you'll win it.

Everyone has ‘that friend’ who makes similar boasts; let’s call ours Dietrich. He’ll boast about how he made a killing on crypto, a fortune on foreign exchange or a surplus on stocks. You take those claims with a pinch of salt. The problem isn’t the facts that you do see – if Dietrich says he made a 76% profit last year, you take his word for it – but the facts that you don’t. You suspect that three other friends, Aanya, Bruno and Chloe, are dabbling in day trading themselves. Yet they’ve never mentioned how they’ve done, probably because they came a cropper. Since only Dietrich is vaunting his investment returns, you have a *selected sample*. This shows why *a fact is not data* : it may not be *representative*. Even if Dietrich’s return claims are true, they’re meaningless, as they don’t tell you how successful day traders are in general.

Before taking the plunge and going into amateur investing yourself, you’d first like to see the profits and losses of *all* your friends. But you can’t force people to disclose their investment record;[2] even if a professor pleaded for volunteers in the name of scientific research, probably only those who struck it rich would come forward. You’d need to be pretty intrepid to get the data you need.

And it took a pretty intrepid person to do so. Terry Odean took an unconventional route to becoming a finance professor. He entered a Benedictine monastery aged fourteen to train as a monk, but dropped out three years later and enrolled in a creative writing degree. He quit that too and took a series of random jobs, including driving a

New York City taxi, before returning to undergraduate studies aged thirty-seven, at the University of California at Berkeley. Terry realized he had a knack for academia and stayed on at Berkeley for a Ph.D. in finance. For his thesis, he wanted to explore whether people's investment performance is as good as they claim – but for that he needed data. The holy grail would be a broker's trading records, because they contain the trades of every client, both winners and losers. However, they're highly confidential, and most people wouldn't dare ask brokers for them. It would be like imploring a hospital to release the medical records of all its patients.

But Terry wasn't 'most people'. He was on a mission. He took every opportunity – tennis matches, parties and random meetings – to beg anyone remotely connected to the brokerage industry to give him the data. Finally, he struck gold, and a large brokerage gave him a database of 78,000 accounts, with client names removed. Importantly, these accounts were randomly chosen by the broker, so Terry had a *representative sample*, not a selected one.

This dataset allowed Terry to write many seminal papers on investor behaviour, mostly with Brad Barber at the University of California at Davis. One influential study calculated the profit that frequent traders make.[†] Importantly, they studied *all* frequent traders in their sample, regardless of whether they struck it rich; they simply picked out every single hyperactive investor without pre-screening for their level of success, and calculated the average return across the group. Recall that Reyes the Entrepreneur claimed to have earned 3.311% in a single day. If he enjoyed that return in each of the 253 trading days in a year, that aggregates to 379,286%.[‡] When Brad and Terry looked at a representative sample of antsy shareholders, it wasn't even close. The average day trader earned a measly 11.4% each year.[3]

The difference between 379,286% and 11.4% is overwhelming. Yet the conclusion might seem underwhelming. No day trader expects to earn 379,286%. Their main hope is to get something positive, to win more than they lose, and 11.4% per year is still comfortably above zero. If you invested for ten years at this average, no better, no worse, your total return would be 194% – you'd triple your money. So Terry spent all that effort, steeled himself to make hundreds of brazen requests, and dealt with rejection after rejection, to no avail. The end result was the same: frequent trading makes money. You won't win as much as 379,286%, but you'll still win big.

But here's the twist. Seeing the full picture involves calculating *two* numbers, and the average profit of restless investors is only the first. The second question to ask is: How much money would you make if you *weren't* a frequent trader? What if you simply invested in the overall stock market, rather than trying to pick a few winners – and then left your portfolio alone instead of chopping and changing it each time news broke out? The alternative –

what would have happened otherwise – is known as the *counterfactual*.

Brad and Terry found that a buy-and-hold investor, who bought the market and didn’t make a single trade, would have earned 17.9% per year. The 11.4% return to frequent trading should be compared not with zero but with 17.9%. This implies a very different, and sobering, conclusion from Reyes and Dietrich. All that time and effort amateur investors put into trading actually worsens their long-term financial security.

The implications are profound. We might think freedom and individual choice are good, as citizens know what’s best for them and shouldn’t be dictated to by governments. But, left to their own devices, they may take actions that hurt themselves. If so, policymakers might design nudges that encourage the public to invest in broad-based funds rather than individual stocks. Sure, you won’t make 379,286% a year – you won’t get rich quick – but you will get rich.

Let’s now highlight each step that Brad and Terry followed, to form a general playbook that we can apply to any question we explore with data, such as what caused Apple’s success. The first step is to state your question in the form of a *hypothesis* about how an *input* affects an *output*. For Brad and Terry, it’s that ‘frequent trading affects returns’.[§]

You then test the hypothesis. For this, you’d ideally like the trading records of every single trigger-happy investor.

That’s impossible, so the second step is to gather a *sample*. What’s critical is that the sample is *representative*, not *selected* – it captures a broad mix of traders rather than pre-screening them on some criterion, such as whether they volunteered to share their record or had an account for five years (both of which would skew the sample to more successful investors). That’s similar to how you’d sample a cake by cutting it vertically so that your slice contains the icing, sponge, filling and base, rather than splitting it horizontally and skimming off only the icing. The extensive compilation of excitable shareholders is known as the *test sample* – you’re testing whether it performs better.[11]

Step three is equally critical – to find a *control sample* that doesn’t have the input. The high returns to fidgety investors might be nothing to do with the input (frequent trading) but just because the market went up. So you need to find out how much was earned by buy-and-hold investors who didn’t trade at all. Step four is to calculate the average *output* across the two samples, which gives you the 11.4% and 17.9%.

You’re tempted to conclude that frequent trading lowers returns, but there’s one final step. Even if frequent trading has no effect on profits, it could still underperform due to luck. In fact, there’s a 50% probability that it does, just like a fair coin with no bias towards heads might still land heads in a single toss. So you can’t just look at whether jumpy shareholders beat or lag the market; you need to take into account two other factors. The first is the *magnitude* of the underperformance. What matters isn't just that 11.4% is lower than 17.9% but that it's 6.5% lower. The second is the *sample size* – how many frequent traders Brad and Terry studied, and for how long. A single punter like Reyes might get lucky over one day, but the researchers analysed 13,293 jittery investors over six years.

We need to work out how likely it is that you get a gap as large as 6.5% from 13,293 investors over six years by happenstance – even if trading frequency didn't affect returns. If the probability is small, then the result is unlikely to have been a product of chance and so it supports your hypothesis. Let's say Brad and Terry crunched the numbers and found that the likelihood was 0.1%.[#] How improbable must a difference be before you interpret it as supporting your hypothesis? It's subjective, but the convention is to have a threshold of 5%, which is known as a 5% *significance level*. Since 0.1% is well below 5%, Brad and Terry's result would be *statistically significant*, allowing them to draw the *conclusion* that frequent trading affects returns.

It seems obvious that the magnitude and sample size matter, and so the need to calculate statistical significance should be clear. If I told you that a coin landed heads more than half the time, you wouldn't immediately label it loaded; you'd want to know how much it beat the midpoint by and over how many tosses. 3 heads out of 5 is barely above halfway (2.5), but 40 heads out of 50 would be striking.

Yet you only need to glance at a LinkedIn feed, Instagram story, or newspaper's daily digest to be flooded with headlines such as 'People who do X are more successful,' 'Companies with Y are more profitable' and 'Countries with Z are happier.' None of these headlines mention how much they outperformed by or for how long, or how many people, companies or countries were studied – yet we lap them up. Even if X, Y and Z were completely irrelevant, there's half a chance they'd still be associated with success.

Despite the power of statistical significance, note that it can never *prove* a hypothesis, so we should be sceptical of claims such as 'indisputable evidence' or 'proof '. Statistical significance simply means that it's unlikely that the result is due to chance – but it's not impossible. On 18 August 1913, at the famous Casino de Monte-Carlo, the roulette wheel landed on black twenty-six times straight, even though the odds are 1 in 66.6 million.[**] Even this highly improbable sequence doesn't prove that the wheel was biased – it could have been fair but just went on a very lucky streak. Or a very unlucky one, depending on where you placed your bets.

The whole playbook – forming a hypothesis, gathering representative test and control samples, testing for statistical significance and only then reaching a conclusion – is the *scientific method*. But Isaacson used a quite different approach. Rather than starting with a hypothesis, he jumped to a conclusion. He identified a number of factors that he claimed were behind Jobs's success, such as his adoption by the right parents, upbringing and focus. Chapter 1 is entitled 'Childhood: abandoned and chosen' and its first section is 'The adoption', so we'll discuss the first factor, but the same concerns apply to the others.

While Isaacson considers only Jobs, others have pointed out how Amazon's Jeff Bezos[4] and Oracle's Larry Ellison were also adopted and deduce that adoption must drive success. Such a conclusion is based on facts. Jobs, Bezos and Ellison were all definitely adopted, and all ended up undeniably successful. However, these facts are meaningless, because *facts are not data*.

You can't just take a selected sample of adopted CEOs who made it big, just like you can't focus only on frequent traders that brag about their winnings. You need a representative sample of dozens – ideally, hundreds – of adopted CEOs, both those who struck gold and those who didn't, and then to calculate their average level of success.[††] And even if most adopted CEOs hit the jackpot, that could be due to an economic upswing rather than adoption. So you also take a control sample of non-adopted CEOs and compute their average success to get the counterfactual. Then you compare the two groups and check for statistical significance.

The scientific method requires you to sample *all* the data. You take a database of CEOs, without pre-screening them for either their success or their adoption status. This database needs to include adopted CEOs who failed and non-adopted CEOs who succeeded, rather than just hand-picking successful adoptees because they fit the picture.

There's nothing wrong with telling entertaining stories if you frame them as just that – stories. You can even conjecture about what led to Jobs's success, as long as you're clear it's speculation. As a biography focusing only on Jobs, Isaacson's book was fascinating and deserved its popularity.

The problem arises when you turn a single anecdote into a general rule. A *Harvard Business Review* article by Isaacson was titled 'The real leadership lessons of Steve Jobs', with the strapline 'Six months after Jobs's death, the author of his best-selling biography identifies the practices that every CEO can try to emulate';[5] *HBR*'s editorial promised readers 'Yes – you, too, can be like Steve Jobs.' You can't change whether you're adopted, so Isaacson instead glorified Jobs's management principles, such as his focus and simplicity, declaring them as a blueprint for all bosses to follow. But Isaacson had no evidence that these principles worked. He didn't study any other CEO besides Jobs who followed them, nor the hundreds of executives who'd reached the top through a different path.

The same problem crops up in our previous examples. We've seen that Belle Gibson's story was false, and stressed the importance of checking the facts. Yet just checking the facts isn't enough. Even if Belle did beat cancer through diet, this would still provide no evidence that diet is an effective treatment. For that, you'd need to study all cancer patients who tried diet as a cure and see how many of them won their fight[‡‡] – but you'll never hear the cases where diet failed, so you have a selected sample.

The biggest problem with Belle's story isn't that it's false. It's that it's only one story. There might be thousands of other stories where clean eating didn't work, but those stories are too ordinary, so they never see the light of day. It's the outlier cases that are new, and so only they make the news.

In *Peak*, Anders Ericsson acknowledges that he had a selected sample. He analysed students who'd already got into Berlin's elite Academy of Music and found that many had practised for 10,000 hours. He didn't demonstrate that practising for 10,000 hours gets you into the academy. 'To show a result like this, I would have needed to put a collection of randomly chosen people through ten thousand hours of deliberate practice on the violin and then see how they turned out.' There may be hundreds of other hopefuls who'd put in 10,000 hours but didn't make it through the doors.

Gladwell writes in *Outliers* that 'The striking thing about Ericsson's study is that he and his colleagues couldn't find any . . . "grinds", people who worked harder than everyone else, yet just didn't have what it takes to break the top ranks.' That's because those 'grinds' didn't get into the academy – so they'd have never appeared in Ericsson's sample to begin with. In the *Fortune* interview he explained how 'The premise of this book is that you can learn a lot more about success by looking around at the successful person.' But this premise is wrong; you need to study failures too.

The narrative fallacy

The idea of forming hypotheses and testing them *before* reaching a conclusion seems elementary – we all know the phrases 'Innocent until proven guilty' and 'Don't jump to conclusions.' However, we forget the basic scientific method when we're told a compelling story.

The formula in Isaacson's, Sinek's and Gladwell's books is a tried-and-tested theme that's behind hundreds of smash hits. Most books have a single big idea to make themselves as memorable as possible, and then find as many examples as they can to illustrate that idea. They only ever consider a selected sample that fits the story, so even if all their facts are correct, they're inconclusive. Bill Joy, Mozart and The Beatles shared many common factors, not just 10,000 hours, so you need to analyse how people *without* the common factor did – like testing sequences without successive even numbers, such as 4–12–26, in the Wason study. Otherwise, you can't attribute their success to years of practice any more than to them all having two legs.

And it goes beyond books. Business school case studies take one company that ended up a storming success or a futile failure and reverse-engineer an explanation for its fortunes – without ever testing whether other companies with the same traits had similar outcomes. As a result, we teach the future captains of industry using stories, not science. If a book, column or case study by a business guru is based on a story, or even several stories, it’s no more reliable than a YouTube video by an amateur.

Outside of business, articles such as ‘How I lost 3 stone in 6 months’, ‘How I got my child into Oxford’ and ‘How waking up at 5 a.m. every day changed my life’ only consider the writer’s personal anecdotes, never a representative sample – let alone a control group. We love to learn from success stories, but you can never identify what drove success unless you also study people with the supposed secret sauce who fell down the mountain and those who reached the summit without it. ‘It worked for me’ doesn’t mean ‘It’ll work for you,’ because a fact is not data.

Why are these stories so compelling? Because they exploit the *narrative fallacy* – our temptation to see two events and believe that one caused the other, even if there were different causes or no cause at all besides luck. Mathematician Nassim Taleb defines it as ‘our limited ability to look at sequences of facts without weaving an explanation into them or, equivalently, forcing a logical link, an arrow of relationships upon them’. We learn that Jobs had a craftsman father and believe this led to Apple’s products being beautifully designed. Or we hear that Apple’s *why* is to challenge the status quo and agree that this must be behind millions of customers buying its products.

The narrative fallacy magnifies the twin biases, and it works both ways. If you’d like people to believe something appealing or extreme but don’t have the data to back it up, just tell a dramatic story; if you’d like your story to go viral, ensure it’s appealing or extreme. The latter is easy, because if you have a single fact to explain, you can reverse-engineer myriad stories that fit that fact and then choose the one that best exploits the biases – just like, if you had a single dot on a graph, you could draw the best-fit line in any direction you want. This flexibility is why we can have two completely different explanations for Apple’s success.

The narrative fallacy is so powerful that it can lead us to believe a story even if it’s false. Jobs debunked the myth that his adoption drove his success: ‘There’s some notion that because I was abandoned, I worked very hard so I could do well and make my parents wish they had me back, or some such nonsense, but that’s ridiculous . . . I have never felt abandoned.’[§§]

The *why* theory also has no backing. Apple never said, ‘Everything we do, we believe in challenging the status quo,’ or anything close, but 26,500 Google hits quote this phrase without checking it because it sounds inspiring. The accompanying biological claims are also untrue. The rational neocortex isn't unique to humans but is shared by many mammals. Sinek's TED talk claims that the limbic brain is 'responsible for all human behaviour, all decision-making', but such black-and-white statements are incorrect when the world is granular. The smell of candied peanuts roasting alongside the River Thames might entice you to hand over £2 even if your neocortex is warning you about their fat and sugar content. However, a £1,000 iPhone purchase is unlikely to be purely spur of the moment but influenced by its features, reviews and recommendations from friends.

Scientific studies have shown how we make up explanations for even random events. Organizational psychologist Barry Staw divided students into groups.[6] He gave each one the same financial data for a company and asked it to estimate that company's future sales and profits. After collecting the forecasts, he told some groups that theirs were accurate but others were wide of the mark. Staw then asked each team to comment on its group dynamics. The high performers said their group was communicative, cohesive, motivated and open to change; the stragglers reported the opposite.

Now that doesn't sound surprising – a better dynamic leads to better performance. But the teams hadn't in fact performed differently. The data was for a fictional company, and Barry told them whether they'd been successful at random. There was zero link between group dynamics and performance. Yet after hearing their outcome, the teams reverse-engineered a narrative to explain it. If a group succeeded, it might say, ‘We allowed for freedom of expression, which harnessed the diversity of our perspectives.’ Upon being told it failed, the same team with the same culture would recount, ‘We had too many cooks and disagreed too much; we should have focused instead on building consensus.’ You can always make up a story to explain success – and people do.

Learning from a blank slate

Growing up in the UK school system, my friends and I were forced to specialize early. Most aspiring doctors study Medicine as their first degree, unlike in the US where it’s a postgraduate qualification. To get into medical school, you need Chemistry A-level. We had to choose our A-levels aged sixteen, making a decision that might affect the rest of our lives.

My school gave us a lengthy psychometric test to find out what careers we might be best suited to. We could then work backwards, like a master chess player who sees ten moves ahead. Once we’d singled out our dream career, we’d figure out the degree we needed to study at university and finally pick the A-levels that would unlock the campus gates.

It seemed like a great idea. Naïve and foolhardy at sixteen, we thought we could plan decades into the future.

We took the test enthusiastically, eager to see what it predicted. But we were disappointed by many of the questions. One asked us to draw a normal ‘S’, then a backwards ‘S’, then a normal ‘S’, and so on, as fast as we could. When we got the test results, we saw that this exercise tried to measure our ‘flexibility’. That sounded like psychobabble – it was ridiculous to make life-changing decisions on how quickly we could alternate ‘S’ shapes. For me at least, the test ended up being a poor predictor. It recommended ten careers, none of which was ‘professor’, or anything related, such as ‘teacher’, ‘author’ or ‘researcher’. Given my confirmation bias, I’ve concluded the test must have been wrong, rather than that I’ve blundered into a career I’m ill-suited to.

As a result of our disappointment, my friends and I took matters into our own hands. We thought there was a much better way to decide our futures: to follow in the footsteps of successful people. We knew better than to read a single biography, so we instead pored over the *Sunday Times* Rich List, which contained the hundred wealthiest people in the UK – being young and foolish, wealth was our only measure of success. We probed into how they found their fortune, paying careful attention to the stepping stones that led to the bounty. Many struck gold through starting their own business, but we were interested in the initial careers that had cultivated these entrepreneurial skills. (As you can guess, none of these hundred tycoons spent any time as an impoverished professor.) And we found out the degrees that launched them into those first jobs.

On the face of it, we made two improvements over Isaacson and Sinek. We had a hundred datapoints, not just a couple of examples. If the problem of the narrative fallacy is that it's based on a single story, can't we solve it by studying a hundred stories? If a single datapoint lets you draw a best-fit line in any direction, surely gathering a hundred datapoints removes such arbitrariness? We also had no pet theory of what drove success, but started from a blank slate. We didn't pick one hero, identify how we thought she reached the top, and then biasedly search for ninety-nine other leaders who'd taken the same route. Instead, we were completely open-minded – we had no preconceptions, and would let the data speak.

Yet none of this mattered, because we still had a *selected sample* : our dataset only contained people that ended up extremely rich. Even if we found a common thread, we'd have had to turn it into a hypothesis, such as 'studying biochemistry leads to success'. But we had no way to test this hypothesis, since we didn't have data on the thousands of other biochemistry graduates who were leading normal lives.[||||] A hundred datapoints don't help if they're all successful people.

Of course, we were immature. Since much of my current work is on the importance of purpose, not financial motivation, I now cringe at the way in which I first tried to choose a career. In our defence, it's not just us. Many gurus purport to base their claims on research, not just case studies and anecdotes, and they equate research to gathering a cornucopia of data. And these gurus include the authors of multimillion-selling books.

Built to Last, by Jim Collins and Jerry Porras, identifies nine principles that seemingly lead to enduringly profitable companies. Chapter 1 ends with over ten pages brow-beating the reader with how much information they gathered, to convince you to put your faith in their principles. They proudly proclaim: ‘we sourced nearly a hundred books and over three thousand individual documents (articles, case studies, archive materials, corporate publications, video footage). As a conservative estimate, we reviewed over sixty thousand pages of material (the actual number is probably closer to a hundred thousand pages). The documents for this project filled three shoulder-height file cabinets, four bookshelves, and twenty megabytes of computer storage space.’ On three separate occasions, they stress how all that number-crunching took six years.

The authors dramatize the research process so that the reader thinks she’s learning something astounding. They compare their journey to Charles Darwin’s five-year voyage aboard the HMS *Beagle*, exploring the Galapagos Islands, likening their findings to his discovery of new species. They’re so taken by their self-comparison to Darwin that, just two pages later, they now say their project took five years. ‘To ensure systematic and comprehensive data collection', they use 'Organization Stream Analysis', which I've never heard of; web searches find no mention of it outside *Built to Last*. In *The Halo Effect*, a critique of *Built to Last* and similar books, business professor Phil Rosenzweig calls this 'the delusion of rigorous research'. It doesn't matter what fancy name you give your techniques or how much data you gather – quantity is no substitute for quality.

What was the problem with the data quality in *Built to Last*? It's our old friend (or enemy): selected samples. The authors found 18 companies which they argued were 'built to last' – successful over a long period, not just a flash in the pan – and then sought 'to identify underlying characteristics that are common to highly visionary companies'. They'd started out with successful companies and then identified a unifying theme (the nine principles): the same methodology my friends and I had used.

To the untrained eye, Collins and Porras went one better than us, as they had a control group. The Rich List only studied success stories, but Collins and Porras also considered failures. They compared the 18 winners with similar companies they deemed not so visionary. Their goal was to identify themes that could explain why Hewlett-Packard was outperforming Texas Instruments, Merck was beating Pfizer, and so on.

But that's not actually a control group. A control group contains hundreds of companies *without the input* – that didn't practise the nine principles. Instead, Collins and

Porras found 18 companies *without the output* – that hadn't been successful. This had barely any effect on their flexibility to come up with whatever explanation they wanted. All it meant was they now needed a common thread that fitted 36 datapoints rather than 18 – that's present in the winners and absent from the losers. If you're asked to explain why one set of countries is more prosperous than another, or one group of schoolfriends gets higher test scores than another, you can similarly take your pick of reasons. Without studying the dozens of other companies that applied the principles but failed, or the hundreds that didn't but succeeded, the authors had no way of knowing whether their explanations were correct.##

This misinference matters. Several of the supposedly visionary companies plummeted shortly after the book's publication, suggesting the magic formula doesn't make you built to last at all – yet millions bought the book hoping to emulate them. The same is true for other books in the same genre which take a set of successful firms and identify a shared theme, such as *In Search of Excellence* and *Good to Great*. None of them has a control group, and the exemplar companies subsequently nosedived. The secret sauce was a scam.

In a nutshell

- *A fact is not data* because it may not be *representative.* Even if a story, case study or anecdote is true, you can’t draw general conclusions, as it could be an anomaly.
- The *narrative fallacy* means we see cause–effect relationships when none exist. We accept story-like reasons for success even if other explanations fit the same facts.
  - To go viral, authors can make up whatever story best exploits the twin biases.
  - People invent reasons to explain their success, even if it’s due to luck.
- The plural of datapoint is not data. Even if an author gives several other examples that share the same narrative, they might be cherry-picked.
  - Ask: Does the author consider cases that don’t fit the narrative – that contain the input but not the output (e.g. an adopted CEO but no success), or the output but not the input (e.g. success without an adopted CEO)?
- The same problem means that you can’t ‘let the data speak’ – gather success stories and identify a common thread – even if you don’t have a preconceived view. Your sample won’t contain cases with the common thread that failed, nor those without the common thread that succeeded.
- Understanding a cause–effect relationship involves the

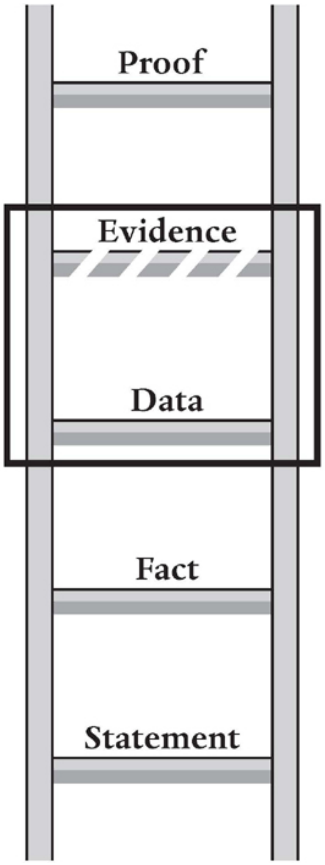
following steps:

- Start with a hypothesis: a given input (adopted CEOs) affects a given output (success).
- Gather a representative *test sample* of cases with the input, irrespective of their output (adopted CEOs, regardless of their success).
- Gather a representative *control sample* of cases without the input, again irrespective of their output (non-adopted CEOs, regardless of their success).
- Calculate the average output across both samples and find the difference.
- Check that the difference is statistically significant – the magnitude and the sample size are large enough that the difference is unlikely to be due to luck.

But even data isn't enough. The world's biggest dataset only takes you an additional rung up the ladder; it doesn't get you to evidence. The next two chapters will explain why.

5

Data is Not Evidence: Data Mining



Few opportunities in life are truly life-changing, but my coffee appointment at the Dorchester Hotel on London’s Park Lane had the potential to be. A few weeks earlier, when cleaning up my junk-email folder, a message caught my eye. Among the notifications of a lavish inheritance, pleas for life-saving donations and promises of unconditional love, one email stood out because of the sender’s name: it was from one of the most famous investors in the world. I braced myself for an investment scam launched under a false identity. Otherwise, what would a powerful money manager want from a lowly assistant professor? But when I opened the email and read its contents, it seemed genuine. We exchanged messages, which gave me further confidence, and when the sender asked to meet in person, I was thrilled.

This was a meeting I didn’t dare be late for, and I arrived nearly half an hour early. I waited in the plush, bright lobby of the £800-a-night hotel[1] where the investor was staying, gazing at the high ceilings and looking through the front window towards the green outline of Hyde Park in the distance. Our meeting time of 10 a.m. came and went, and the doubts started to creep in. The emails appeared authentic, and the sender seemed to know my research. Perhaps this was naïve acceptance on my part – I wanted the messages to be genuine. My rational System 2 kicked in for the first time since receiving that email, and I wondered if I’d been duped.

Ten minutes later, the investor arrived, looking the very image of a master of the universe – radiating both authority and calm. The financier, whom I'll call Xinyi, wished to launch a fund that backed pro-diversity companies, and wanted evidence to buttress her strategy. She'd heard of my research showing how the 100 Best Companies to Work for in America – firms that go above and beyond in how they treat their employees – beat their peers, and hoped this might be what she needed.[2] I explained that the Best Company assessment does take diversity into account. Yet it's far more than that, and it's impossible to know whether it was diversity, or the other aspects of being a Best Company, that drove the outperformance I'd found.[*] Employee satisfaction is granular, and diversity is only one item under the umbrella.

Undeterred, Xinyi asked if I could adapt my methodology to conduct a new study, focused specifically on diversity. I said I could, by replacing Best Company status with a diversity measure, and gave examples of the many ones available. Xinyi was enthused, and asked if I could perform the analysis. If it worked out, I could partner with her in the launch of this new fund.

I stepped out of that lobby walking on air. This was a golden opportunity, and if the results panned out, the benefits would be endless. I'd land a top publication, which would be highly cited, since diversity was a hot topic. Beyond academia, I'd work with one of the leading investors on the globe, who'd catapult me out of the ivory tower on to the main stage. I could be heralded as a champion for diversity, in turn opening many other doors. Quite apart from the instrumental pay-offs, diversity was something I was intrinsically passionate about. I'd always been one of the few non-white faces at school, at university, in investment banking, and even in my leisure time – such as on the football terraces where I had a season ticket in the late 1990s.

Given the many diversity measures available and thus analyses to run, I approached one of my Wharton MBA students to work with me, whom I'll call Dave – because that's his name. He had strong quantitative skills, having scored an A+ in my class, and a passion for ethical investing. We studied a Thomson Reuters database, which provides data across eighteen different dimensions of corporate responsibility, one of which was diversity. Within diversity, there were dozens of areas. Xinyi was particularly interested in gender, and we found 24 relevant measures.[†] We crunched the numbers 24 times, hoping to strike gold.

But instead we struck mud. Out of those 24 measures, 22 were *negatively* associated with company performance. For some of those 22, such as the percentage of female board directors, the relationships were statistically insignificant – sufficiently weak that they could have been due to chance. For others, like having a maternity-leave policy, the link was both negative and significant. Out of the two bright spots, the percentage of women managers had a positive link, but it wasn't significant. However, there was one association that was significant in the direction we wanted – the number of diversity controversies reported in the media. The fewer the headlines, the stronger the performance.

It was clear what we should do if we wanted to work with Xinyi – report only the one positive and significant result. Or we might disclose both positive results, to give the impression of honesty, as we could concede that one was insignificant but we were reporting it for transparency. But my job as a professor was to do scientific research. Even if I only used the study to launch a fund and gave up on the idea of publishing an academic paper, I'd still be spreading misinformation. Research is research, regardless of what it's used for, and scientists can't just pick and choose the results they want.

We emailed all 24 results to Xinyi, who was disappointed but graciously thanked us for our efforts. Dave wrote up the results into a thesis for his MBA; I went back to my other projects, thinking this was the last I'd hear of it. Despite my hopes after the initial meeting with Xinyi, I took this failure on the chin. Over my short career, I'd already tested several hypotheses that didn't work out, so I knew that disappointment was simply part of the research process.

Six months later, a news article grabbed my attention. Xinyi was launching a fund based on the premise that female-friendly companies perform better – the exact thesis our analysis had contradicted. She quoted research from a company I'll call Fixit that claimed to find a huge effect of diversity on performance. Fixit used a diversity metric that was none of the 24 that Dave and I studied, and entirely different performance measures. Not surprisingly, Xinyi made no mention of the tests that Dave and I ran for her and which didn't pan out.

This episode highlights a key reason why *data is not evidence*: it may be the result of data mining. Data mining is where researchers engage in a biased search for a particular conclusion – they conduct hundreds of different tests, hide those that don't work and jump on the one that hits the bullseye. As a result, there's a simple path to launching an influential paper – mine the data, hope and pray you'll find something significant, and report only that result.

In fact, you don't even need to hope and pray. You just need to run enough tests. Even if there's no true link between the input and output, one might arise in the data due to luck. If you toss a fair coin enough times, there will be streaks of six heads;[‡] if you test a hypothesis 100 ways, 5 of them will be significant at the 5% level, even if the hypothesis is false.[§] This means, on average, you only need to try 20 times to get what you want.[||] 'If at first you don't succeed, try, try again' isn't just an abstract proverb – it's true in real life when it comes to data mining.

How can you get enough tries to ensure you succeed? By experimenting with different measures. Both Xinyi and Fixit are guilty of data mining in this example – Xinyi hand-picked the Fixit research because it claimed what she wanted, and Fixit likely knew their study would be more impactful if it found significant results – but we'll refer to Fixit, as they actually crunched the numbers. Starting with the input, Fixit could have studied any one of the 24 diversity metrics in Thomson Reuters, plus the dozens, potentially hundreds, in other databases.[#] They stumbled on one that gave them what they wanted – comparing companies with three or more female directors against those with zero.

Fixit also played around with the output: financial performance. There's one indicator that's head and shoulders above the rest – shareholder returns.[**] That's how much shareholders get from investing in the company, and so it was the only yardstick Dave and I studied. But Fixit used the profit margin instead.[††] The profit margin misses out dozens of other drivers of shareholder returns, such as future prospects, new product launches and management changes. Yet Fixit chose the profit margin because it worked. It was the one output that, by chance, happened to be correlated with their measure of diversity. Fixit's results gave Xinyi what she needed to launch her fund, attracting millions from eager investors who lapped up her claims.

Mining fools' gold

But is data mining a problem? If companies with at least three women on the board outperform those with zero, that's a fact. Actually, it's more than a fact; it's data – the link is there in cold, hard numbers. Even if other measures of diversity don't work, that doesn't change the reality that this one does.

It *is* a problem, for two reasons. First, Xinyi wanted to launch a fund based on gender diversity in general. Her prospectus highlighted how she'd assess a company's diversity – not just whether it has three or more women directors but the gender of the CEO, female-friendly working practices, and so on. Yet the data didn't show a link between any of these other measures and performance. You can't extrapolate from a tree to the forest when the world is granular.

Couldn't Xinyi instead launch a fund focused only on the number of female directors which *is* correlated with performance according to Fixit? Unfortunately not, because of the second problem – even a significant result might be due to chance. Statistical significance doesn't prove a relationship exists; it only shows there's less than a 5% probability that there's no actual link and you just got lucky.[‡‡] If there's no true association with performance, the investment strategy won't outperform in the future. Indeed, Xinyi's fund has underperformed both the US and world markets since its launch.

This is the problem of *spurious correlations*. In Chapter 4, we highlighted how statistical significance gives us data, not just facts. But it doesn't take us to evidence – if a relationship is due to randomness, it doesn't support any hypothesis, which is the hallmark of evidence. Author Tyler Vigen has a website listing some of the most bizarre correlations to underscore this point. The number of people killed by venomous spiders in the US is correlated with the number of letters in the winning word of that year's National Spelling Bee; murders by steam, vapours and hot objects are linked to the age of the reigning Miss America. Moving to less morbid outcomes, per capita consumption of margarine is associated with the divorce rate in Maine.

The worrying implication is that misinformation is hard to police. You'd hope that regulators would crack down on funds making misleading claims, but nothing in Xinyi's prospectus was incorrect. Every statement was backed up by data, but that data didn't amount to evidence because Fixit mined it. Lawmakers have no way of knowing how many other results a company has hidden – so even if they've signed off on a prospectus, it may still contain lies. Since we can't rely on regulation to protect us, we need to learn how to take responsibility ourselves.

Closing down the mine

So how can we guard against data mining? The first step is to ask: *Are the input and output being measured in the most natural way?* For fund performance, any investor knows that returns are the only possible yardstick, but they might forget this if they like the profit margin result. In other cases, there might be several equally valid measures. Even if there's nothing wrong with the indicator being used, ask yourself if others exist. The number of female directors is a perfectly reasonable way to gauge diversity, but so is the fraction of new women employees. If so, does the study check for *robustness* – show that its results continue to hold when using the most plausible alternatives?

But the problem of data mining is much bigger than that. Xinyi wished to launch a fund based on one specific investment idea, gender diversity, yet still had a wide range of measures to choose from. What if Xinyi simply wanted to launch a fund and didn't care about the strategy – whether it be diversity, clean energy, CEO quality or something else – as long as it works? The world is then her oyster. She can try correlating stock performance with thousands of different inputs, ranging from the sensible (the CEO's education) to the absurd (the CEO's shoe size), and almost certainly she'll turn up a couple of significant ones. The problem isn't freedom to choose how to measure an input – there's only one way to gauge shoe size – but freedom to choose that input.

To solve this issue, we ask a second question: *Is it plausible that the input affects the output?* Answering it requires not statistics but *common sense*, a theme we'll come back to time and again in this book. Recall that the scientific method entails *first* coming up with a hypothesis and *then* testing it. If there's a common-sense explanation for why the input and output are linked, such as CEO education affecting firm performance, then it's reasonable to believe this pair was the only one the researchers studied. However, if there's no logical association, the result was probably data-mined. There's no clear reason why hydrogenated vegetable oil consumption sparks marital discord, so the link between margarine and divorce is likely random.

Yet this second question isn't foolproof, because researchers can mine their hypothesis, not just their data. They can run hundreds of tests, see what works and make up a hypothesis after the fact. If they found that CEOs who like red perform better, they can then bury their heads in the psychology literature, hoping to find evidence that red triggers dominance and so enhances performance – and indeed such a study exists.[3] The authors could write a paper starting with the hypothesis that red-loving CEOs will conquer all, using this article as justification, and then testing it; you'd be none the wiser that things happened in the opposite order. Or, if the data showed that CEOs who like red perform worse, the authors could hunt for research that seeing the colour red leads to fear of failure – and indeed such a study exists.[4]

How can we tell if a hypothesis is reverse-engineered? To see how, we'll delve into my past to look at the first paper I ever published, where I had to rebut concerns of data mining. I started it at the beginning of the second year of my Ph.D. But we need to first go a couple of months further back, to the summer of 2004, when the seeds of the idea were sown.

Football frenzies

‘The trade I’m suggesting is a 2y forward 2s/10s bull steepener. It’s elementary to execute. We simply take the DV01s into account to make the trade duration-neutral, by choosing the notionals so that the product of the notionals and the durations are the same on both legs of the trade.’

As I was trying to get my head around the not so elementary idea Mike had just described, Julia, several desks over, started screaming: ‘I need a million Loonie, a million Kiwi, a million Stoki, a million Noki!’

It was the first summer of my Ph.D. Most students had stayed at MIT to work on their research, but I was back at Morgan Stanley. I’d previously been in their investment banking division in London, but switched to sales and trading in New York to taste another side of real-world finance.

The trading floor was crazy. For most people, the constant noise would make it impossible to concentrate, but I loved the energy. Julia, an executive director in foreign exchange sales, would make a daily request for the latest currency rates. A trader in another part of the room would then shout back the current price of the Canadian dollar, nicknamed Loonie because the CAN$1 coin contains a picture of a loon, a native Canadian bird. Then I'd hear an Antipodean accent yell the price of the New Zealand dollar (Kiwi), followed by a Scandinavian voice scream the latest rates for the Swedish krona (Stoki) and Norwegian krone (Noki).

Loud as they were, the traders at least hollered in an orderly manner, one after the other. But if news on the economy broke out, such as unemployment numbers, you'd hear curses or cheers in surround sound. At times, the atmosphere approached that of a football match. And a few weeks into my internship, it *became* that of a football match, because the 2004 European Championships kicked off.

I played 'soccer' at MIT, and when I heard someone from the stands shout, 'Dude, that guy just totally headbutted the ball!', presumably because he'd never seen a header before, I took it as a sign that Americans didn't know or care much about football. Perhaps they might be interested in the World Cup but surely not the Euros? I quickly learned I was wrong. Many born-and-bred Americans had European heritage, as shown by their surnames, and cheered for the country of their ancestors even if they'd never set foot on its shores.

This was the year Greece was the surprise package, winning the tournament despite being 150–1 outsiders at the start. On the way, they caused many upsets – and upset was the emotion aroused in my colleagues. Powerhouses Germany, Spain and Italy were knocked out in the first round, sparking anger that not even the most frightening unemployment number could induce. These were managing directors and executive directors used to winning or losing hundreds of millions of dollars in a day, yet a market rout didn’t incite nearly the same deskbashing, Neanderthal screams and impassioned language as a football defeat. After defending champions France were knocked out in the quarter-finals by Greece, an English guy on my desk taunted a French senior trader with ‘I’m going to have some Greece-y chips now.’ The Frenchman stormed off and didn’t return to work for several days.

When summer was over, I returned to the calmer halls of MIT for my second year, which kicked off with a course on how to do research with financial data. The professor, Tim Johnson, wanted to get us into good habits and so his first lecture taught us about the dangers of data mining. He introduced us to the literature on how sentiment affects the stock market. This research aimed to show that investors are irrational – they don’t base their trades on only sensible factors like profits and inflation but are swayed by their mood.

Since you can’t observe the sentiment of individual traders, researchers study the sentiment of a country. To do this, they need a measure that affects the nation’s mood but not its economy. Finding one is difficult, because many factors that influence the former will also drive the latter – a plane crash hits the travel sector, an election affects taxes and a pandemic shuts down the economy. If these events cause the market to tumble, that could be entirely rational and nothing to do with sentiment.

Prior papers had come up with clock changes,[5] seasonal affective disorder,[6] weather[7] and lunar cycles[8] as inputs that affect feelings but not finances. They'd all found significant results and been published in top scientific journals, but Tim remarked that they were still controversial due to concerns about spurious correlations. Is the stock market really affected by whether the moon is in a waning crescent or a waxing gibbous phase?

Tim encouraged us to take a step back and look at the bigger picture. If you wanted an event that shocks a nation's mood, would clock changes, for example, be the first thing you'd look at? There's certainly a story for why clock changes might matter: they disrupt traders' sleep patterns, making them grumpy and more likely to sell. However, it's not clear how plausible this mechanism is. Traders are pretty good at getting by on little sleep – the roughest night can be undone by the strongest coffee. Any remaining effect on emotions is so tenuous that it's hard to convince readers that clock changes are the first and only event that you studied. To be immune to criticisms of data mining, Tim explained that you needed such a strong measure of mood that even the harshest sceptic would accept it was the sole one you examined.

That summer immediately came back to me. I recalled how the senior traders, who were sharp as needles after a sleepless night and unfazed by whether thunder or sunshine was ruling the skies, sank into despair when their team lost a match. The effects of Euro elimination couldn’t be undone by the freshest grande, quad, one-pump, no-whip, oat-milk mocha – not even one with a caramel drizzle on top. Football results seemed a much bigger driver of mood than anything else I’d seen on the trading floor.

As I looked into the research on how sports affect emotions, I learned that the impact of Euro 2004 on Morgan Stanley traders wasn’t an isolated example; the effects are widespread. When England lost to Argentina in the 1998 World Cup (on penalties, of course), heart attacks shot up over the next few days.[9] Across the pond, suicides rise in Canada when the Montreal Canadiens are eliminated from the ice hockey Stanley Cup.[10] While Canadians kill themselves, Americans kill each other – when a team is knocked out of the NFL[§§] play-offs, murders go up in the local city.[11]

These effects are much stronger than for the previous mood measures – I don’t know of anyone breaking the sixth commandment because the clocks sprang forward. What’s more, losing a game doesn’t just make you depressed about your team; the effects spill over into general life. When the Ohio State American football team wins, lottery purchases go up in Ohio because people feel more optimistic. So it’s quite plausible that sports matches have a strong enough impact on mood to feed through to the stock market.

I crunched the numbers and found that, when a country loses a major international game, its stock market falls significantly the next day. Diego García and Øyvind Norli, two professors at Dartmouth College, happened to be working on the same idea with identical results, so we joined forces. Together, we compiled data on 1,100 games across the World Cup, European Championship, Copa America and Asian Cup and found the results held at large scale.[12] To put some numbers on it, a World Cup loss is associated with the national market declining by 0.5%. Applied to the UK stock market, that's about £10 billion wiped off in a single day – just because England can't take penalty kicks.

This paper was the first chapter in my Ph.D. thesis at MIT and published in the *Journal of Finance*. You might think it crazy that MIT awarded a doctorate for a study on football, and that it came out in a top academic outlet. But the beauty of the beautiful game is that it satisfies the twin requirements of having little effect on the economy yet a large effect on sentiment. This allowed us to convince the MIT thesis committee and the *Journal of Finance* editors that it was the only measure of mood that we looked at, and that we hadn't data-mined.

This principle applies far beyond our paper. As a researcher, the best defence against data mining is a strong hypothesis. As a reader, if you suspect a correlation might be spurious, ask yourself – *what's the question the authors are exploring, and is the input the most logical way to do so?* If the topic is whether sentiment affects the stock market, ask whether clock changes would be the first measure of sentiment you'd think of. If it's whether CEO characteristics predict company performance, would her favourite colour be the number-one trait you'd investigate?

But even if you're willing to believe that an input was the only one studied, and that it's measured in the most plausible way, a master miner still has two more tools in his kit. Let's now look at the first.

Turning the tables

The hard slog was nearly over. A first-year investment banking analyst, I'd finished running the numbers for the deal we were proposing to our client. I'd studied various financing scenarios, written up the strategic rationale for the transaction and finalized the section on risks and mitigants – incorporating the numerous, and sometimes contradictory, comments from hordes of senior bankers along the way.

There was one section still to be written – 'Morgan Stanley credentials'. While the rest of the pitchbook sought to convince the client to do the deal, this final flourish would persuade them to do it with *us*. For that, I needed to construct league tables showing the value of transactions we'd completed versus our competitors. Since we were pitching a Chemicals deal to a German client, I compiled statistics for European Chemicals activity over the last three years. The numbers came out pretty favourably, with Morgan Stanley ranked at number three.

But a bronze medal wasn’t enough for my boss, Mark. He asked me to run the numbers again. What if we looked at the past four years, or the previous two? Perhaps we could restrict the criteria to include only deals worth €100 million or more? And why European Chemicals? That was the name of Mark’s team, but given the client’s location, couldn’t we try German Chemicals? Or maybe Global Chemicals, to stress how Morgan Stanley offers worldwide expertise? Surely there was some combination which would give us a higher place on the podium?

I went back to the drawing board and tried some other permutations. Cutting out deals below €100 million actually saw us drop to fourth. I wasn’t deterred, as I still had other dials to turn. After further crunching, I stumbled upon a combination that improved our ranking to runner-up. I didn’t bother showing this to Mark, as I knew he’d only be satisfied with first place. I spun the wheels several more times, and at long last I found the magic set of ingredients that saw Morgan Stanley take its rightful place at the top of the tree.

This example shows just how pervasive the problem of data mining is. The sentiment studies illustrated the freedom we have to choose our input – clock changes, weather or football results. My work with Xinyi demonstrated how, even if you’ve nailed down one input, there might be dozens of ways to measure it. With the league tables, we had no flexibility on either our input (we were pitching for a deal, so we had to show our deal experience) or how to measure it (the value of a deal is unambiguous).

But we did have latitude on our sample: what deals to include. We knew the result we wanted and could cherry-pick our ingredients so that ‘computer says yes’. This type of data mining is known as *sample mining* – choosing not your input, output or measures but your sample.

Sample mining doesn’t just occur when there’s a multimillion-dollar-deal fee at stake but also when your reputation is on the line. When we were investigating diversity for Xinyi, Dave and I were surprised to find no link between female board members and performance, because Thomson Reuters – whose data we were using – claimed that ‘companies with no women on their boards on average underperformed relative to mixed boards’.[13] The fine print revealed that they had only studied 2007 onwards, which seemed odd, since their own data was available from 2002.

We reran our numbers, using the same time period as Thomson Reuters, and sure enough we also found a positive result. But when we considered the full sample, the relationship became (insignificantly) negative. That result wouldn’t have been great for Thomson Reuters’ PR, so they got a different one by throwing away five years of data – nearly half their sample.[‖‖] Thomson Reuters titled their study ‘Mining the metrics of board diversity’, rather ironically, since they’d engaged in a brazen case of sample mining.

Defending against sample mining

How can you tell if researchers have sample-mined? By checking if they've run *out-of-sample* tests – shown that their results hold in a different sample, such as a new time period or a different country. However, you might worry that they tried dozens of alternatives and only reported the ones that worked. So the best approach is for someone else to choose the sample for them, to prevent such hand-picking.

For investment banks, this 'someone' is a potential client. If you're poised to pay a multimillion-dollar fee, you can ask for whatever you want – including league tables under your chosen criteria. What about a scientific study? Academics (unfortunately) don't have paying clients, but the public is our client and prominent members of the public can ask for out-of-sample tests.

On the opening day of the 2014 World Cup, CNN International news anchor Richard Quest interviewed me on his show, *Quest Means Business*. I'd already done several interviews on my football paper, so I wasn't too daunted. But Richard wasn't shy about vocalizing his incredulity that sports could affect stocks. Midway through my first answer, he interrupted me with 'Oh come on! How? Why?', and continued to grill me throughout. I kept responding with evidence, and Richard open-mindedly changed his view.

Once the cameras stopped rolling, he remarked that it was a fascinating study and thanked me for coming on. Just as I was about to relax, he asked me to track every result of the current World Cup to see if my theory held up and email him the results. This was the first time an anchor had put me on the spot like this. It crossed my mind to conveniently forget his request, but Richard might infer from my silence that things didn't work out. Plus, I was interested to see for myself.

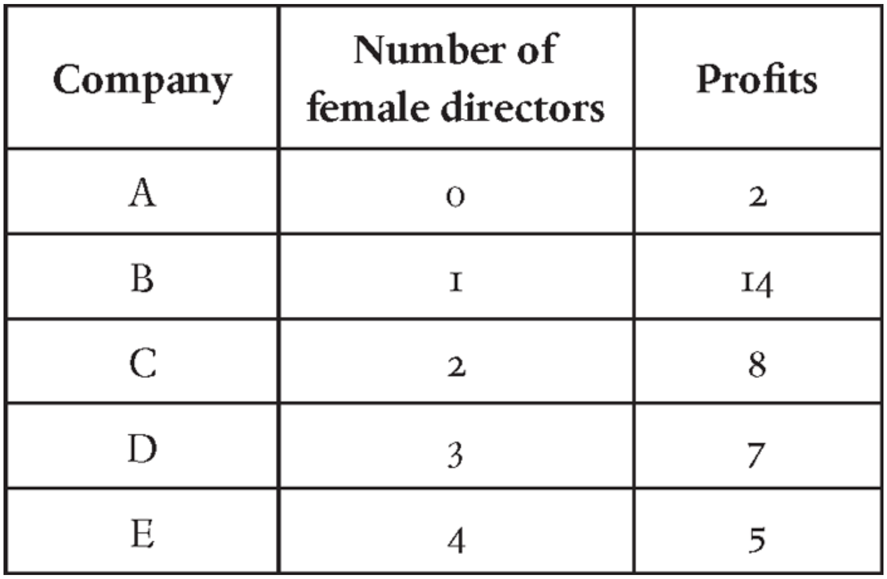
Over the course of the tournament, two thirds of defeats ended up being followed by a decline in the losing team's national stock market (relative to the world market) – significantly higher than the one half you'd expect if it were random. On average, a loss was followed by underperformance of 0.2%; a defeat for the 'big seven' football countries (England, France, Germany, Italy, Spain, Argentina, Brazil) was followed by a drop of 0.4%. Richard interviewed me twice more during the tournament, and I explained how the results were panning out, with each conversation as animated as our first.[14]

Most studies don't end up on CNN, so there's no Richard Quest to grill researchers on the public's behalf. As I'll discuss in Chapter 9, scientific journals peer-review papers before publishing them, and the reviewer plays gatekeeper. For the football study, ours asked us to examine other major international sports beyond football.## We explored cricket, rugby, ice hockey and basketball and ran the tests with bated breath – if they didn't work out, our paper would be finished, as it would suggest we'd got lucky with football. But we found that losses in rugby, cricket and basketball also led to market declines, with only ice hockey bucking the trend. Even including ice hockey, the average loss effect across the four new sports was statistically significant. Our results held up out-of-sample.

The three categories of data mining we’ve seen so far – choosing your inputs and outputs, how to measure them, and your sample – are all about rigging the deck in your favour. What about if you have to play the hand that you’re dealt: you’re given your cards and aren’t allowed a reshuffle? You still have one final trick up your sleeve, which we’ll now reveal.

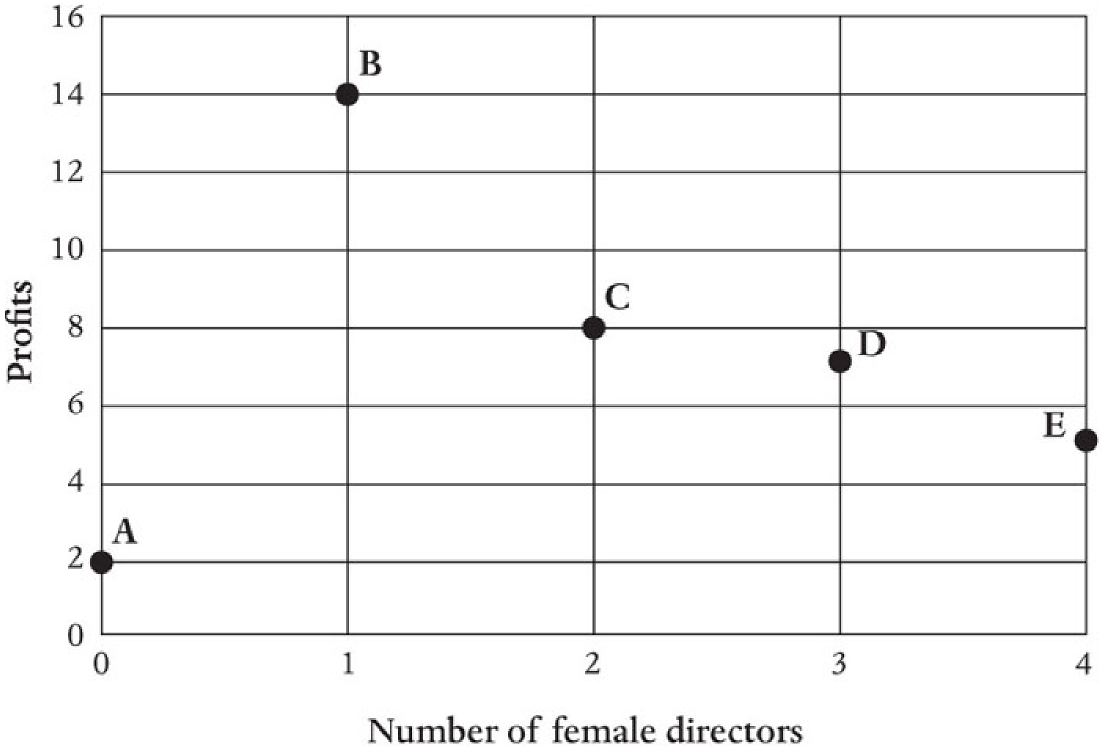
Cooking the books

Fixit obtained significant results by comparing companies with three or more female directors to those with none. But Dave and I also studied board diversity – it was one of the 24 measures we investigated – yet we found no link. To see how we reached different conclusions, consider the following hypothetical data:



Company	Number of female directors	Profits
A	0	2
B	1	14
C	2	8
D	3	7
E	4	5

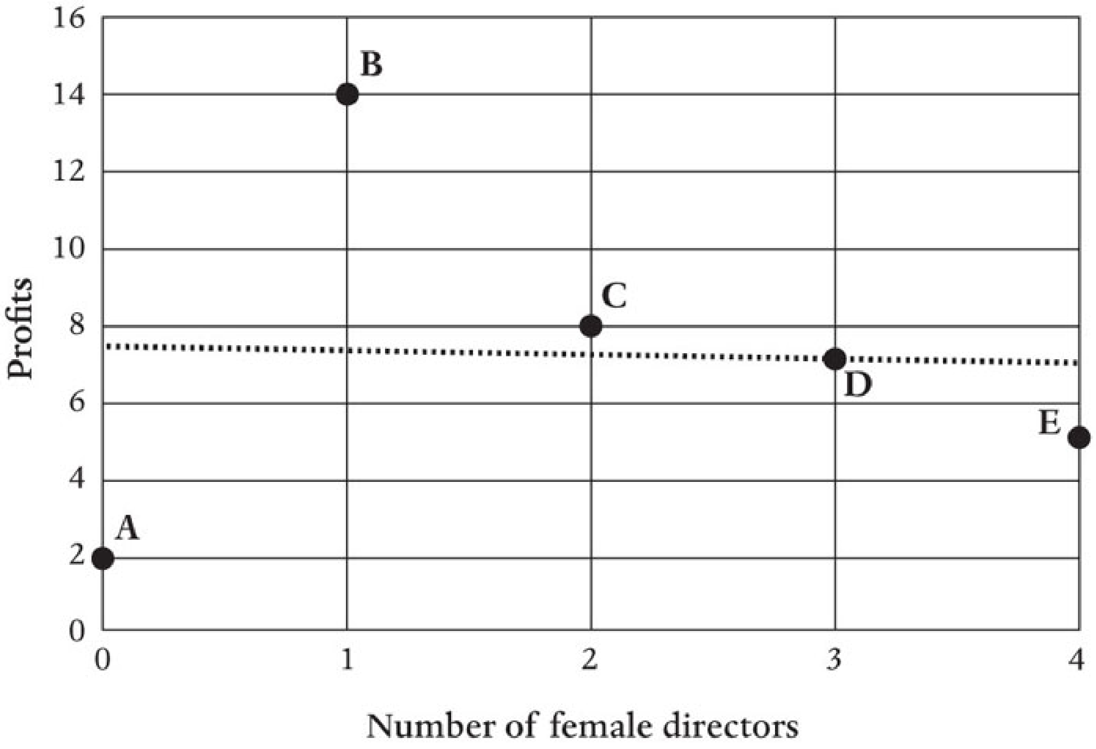
How would you assess whether the number of female directors is associated with firm performance? The most natural way – and indeed the correct one – is to plot a graph of one against the other:



To find the relationship between the datapoints, you draw a *best-fit line*. The slope of the best-fit line gives you the link between board diversity and profits – if a company has one additional female director, how much higher do its profits tend to be, on average? If you consider only B, C, D and E, the line would slope downwards. A throws a spanner in the works; perhaps its presence means that the best-fit line should trend slightly up?

Fortunately, we don't need to rely on guesswork. A *regression* calculates what the slope of the best-fit line should be, by finding the line with the lowest average gap to each of the five datapoints.*** You plug all ten numbers from the table into a formula, and it gives you the slope, also known as the gradient.

If you run a regression on the above data, you get a gradient of –0.1, which tells you that the best-fit line should be slightly downward-sloping:



The gradient of –0.1 means that, if you take two companies, F and G, and G has one more female director than F, its profits will be 0.1 lower, on average. Regression is the method Dave and I used, which is why we found a slightly negative relationship. After calculating the slope, you can then test it for statistical significance. We found that –0.1 was insignificant – so small that it’s probably due to randomness rather than diversity actually harming profits.

Fixit did something different. Instead of running a regression, which uses all the data, they needed an excuse to get rid of that pesky company B, which had stellar performance despite low diversity; C was also pretty inconvenient for the result they wanted. So they used the final weapon in the miner's arsenal: *grouping*.[†††] They divided the data into buckets – one containing those with three or more female directors (D and E) and another consisting of those with zero (A). This allowed them to ignore the meddlesome B and C, because they didn't fit into either category. Since the firms in the three or more bucket (D or E) had average performance of (7 + 5) / 2 = 6, and the all-male company (A) yielded only 2, Fixit were able to claim that diversity improves performance by an impressive 4.

Most of this chapter has shown how researchers can datamine by selecting their ingredients – the input, the output and the sample. Their final lifeline is that they can cook whatever dish they want with those ingredients, and throw away any they don't like the taste of.

Grouping also solves a second problem for Fixit. If you compare D and E, the latter has one extra female director (4 vs 3) but worse performance (5 vs 7), which went against Fixit's desire to show that diversity boosts profits. They might want to conceal E's greater diversity, and grouping achieves this by putting D and E in the same bucket. Both count as having at least three female directors; the precise number no longer matters. Grouping converts the shades of grey in the different numbers of female directors (0, 1, 2, 3 or 4) into black and white. All that counts is whether you have 0 or at least 3; it's irrelevant whether it's 3, 4, or 34.

If comparing firms with at least three female directors to those with zero didn't work, Fixit could have tried a different recipe. They could have contrasted those with at least three to those with below three – defined the *control group* of non-diverse companies as having two or fewer women, rather than zero. And if that also failed, they could have experimented with one or fewer. Similarly, Fixit might have specified the *test group* of diverse companies as those with at least two female directors rather than three – and if this was to no avail, they could have plumped for at least four.

We've just slipped in the terms 'test group' and 'control group'. But Chapter 4 said that the correct way to test a hypothesis was to compare a test group with a control group, so why am I whinging about this now?

Grouping makes sense when the input is *binary* – when it can only be zero or one, or black or white. Your CEO is either adopted or she isn't, so you compare firms with adopted CEOs to those without. However, many inputs aren't zero–one; they're *continuous*. Boards aren't either diverse or non-diverse; they have different levels of diversity – four female directors is greater than three, and three is more than two. As a result, the data isn't neatly divided into test and control groups, so you can't do a simple binary comparison.

A regression uses *all* the data and takes into account a company's actual level of diversity, undistorted by any grouping. The only answer that you can get is –0.1; you have no latitude to mine the data for a different outcome.[‡‡‡] In contrast, grouping gives you tremendous flexibility because you can hand-pick your groups. Whenever we see data divided into two buckets, ask: *Is the input naturally binary, or naturally continuous and the researchers imposed the grouping?* If the latter, the authors may have chosen whatever groups gave the conclusion they wanted. We need to check whether they've shown that the results continue to hold under a standard regression.

How can people get away with ignoring the shades of grey? Because of black-and-white thinking. We like to divide the world into good and bad. 'Diverse companies beat non-diverse ones by 4' has a clear punchline. With a regression, there's no such thing as diverse or non-diverse; instead, you summarize it with the slope of the best-fit line. A slope of –0.1 means that one additional female director is associated with lower profits of 0.1, on average. This still has a real-world meaning but is less catchy than the goodies beating the baddies by 4.[§§§]

In a nutshell

- *Data is not evidence* because it may be the result of *data mining*. The researchers may have run dozens of other tests that failed and only reported the ones that worked. Even if a relationship is statistically significant, it may still be due to luck.

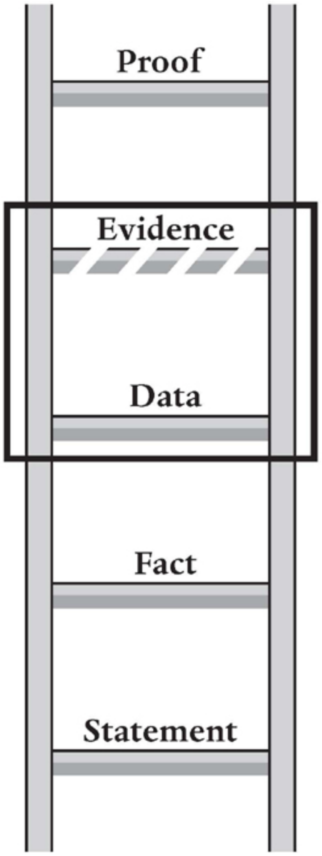
- They might have tried different measures for their input and output, such as different metrics for diversity.
    - Ask: Are the input and output measured in the most natural way? If not, does the study demonstrate robustness to alternative measures?
- They might have tried different inputs. To predict stock returns, researchers can use anything from the CEO’s education to her shoe size. If they run enough tests, some will be significant, even if there’s no actual relationship: a *spurious correlation*.
    - Ask: Is it plausible that the input affects the output? What’s the question the authors are exploring, and is the input the most logical way to do so?
- They might have cherry-picked their *sample*. This can include the start and end dates, or the criteria to get into the sample (e.g. deals over €100 million).
    - Ask: Have the authors conducted out-of-sample tests that use a different sample?
- Even if researchers have to play the hand they’re dealt, they can data-mine by *grouping*. They might compare three or more to zero, three or more to two or less, or two or more to one or less.
    - Ask: Is the input *continuous*? If so, and the

researchers have made it *binary* by grouping, check if the results still hold up in a regression.

But even if you have a rock-solid hypothesis and have shown your results are robust to alternative measures and out-of-sample tests, there’s a second reason why data is not evidence. That’s the subject of the next chapter.

6

Data is Not Evidence: Causation



Data and equations had always been the focus of my life, until my son Caspar was born. Some moments remain as vivid as if they were yesterday – seeing Caspar resting on his mother immediately after the delivery, hearing his first sneeze, and marvelling at him opening his eyes to the world for the first time. Others are blurred, and I remember very little between the epidural going in safely and holding my wife's hand as she started to push.

Those hazy memories are partly because Caspar was thankfully born without incident. But as a friend, and father of three, wrote to me after I shared the news, 'Now it's over, and now it starts.' Taking after his dad, Caspar was almost permanently hungry, with a calm visage transforming into a desperate cry at the flick of a switch. My wife's milk supply was still developing, so even after Caspar had drained her dry, he still wanted more. We faced the burning question that any new parent has to grapple with – do we turn to the bottle?

There's apparently ironclad evidence that breastfeeding has a whole range of mental and physical benefits for both baby and mother. We'd taken a National Childbirth Trust antenatal course that included two classes dedicated to breastfeeding, where the instructor went through a 114-page PowerPoint deck brimming with information. The bulk was on how to breastfeed, but it started with the benefits – breastfeeding is tough, so knowing the payoffs would encourage bleary-eyed parents to persevere. To my delight, the instructor included full academic references in the footnotes, so we had data, not just statements.

And it’s not just the NCT: almost everyone tells you ‘Breast is best,’ and similar advice abounds on the internet – importantly, from reliable sources. Being a bookworm, one outcome I was particularly interested in was child IQ. Googling ‘breastfeeding IQ’ gave the first hit as BBC News; the Google preview read ‘A long-term study has pointed to a link between breastfeeding and intelligence. The research in Brazil traced nearly 3,500 babies and found those who had been breast-fed for longer went on to score higher on IQ tests as adults.’ The second link was to a study in *BMC Pregnancy and Childbirth,*[1] with the excerpt ‘Breastfeeding was positively associated with full IQ at 8 years and negatively associated with hyperactivity/attention deficit at 4.’ The third referenced a paper published in *Frontiers in Nutrition,*[2] with the extract ‘Breastfeeding was positively associated with IQ performance in children and adolescents. On an average, more breast-fed participants had high IQ.’

These studies confirmed everything I thought to be true. Natural breast milk is surely superior to formula fabricated by a corporation. Companies have thousands of scientists passionate about newborn nutrition, but they’re no match for the evolution of billions of humans over millennia – just as synovial fluid, which greases our joints, is forty times more slippery than any synthetic lubricant.[3] It seemed a clear-cut case: reaching for the bottle to get short-term peace would be at the expense of Caspar’s long-term development.

But among the hazy memories and overwhelming moments from the birth, I remembered one comment from the obstetrician. Because Caspar was born early, he was underweight, so she recommended we top him up with formula if the need arose. Not only would that boost his growth, but the liquids would also help flush out bilirubin, a by-product of breaking down red blood cells that’s particularly elevated in babies born early. High bilirubin levels cause jaundice, which, left untreated, can lead to brain damage and even death. Thus, formula might have benefits as well as costs, particularly in Caspar’s case. The issue might not be so black and white.

The NCT deck ended with their breastfeeding support number. I called up and explained our dilemma. Perhaps the relationship between breastfeeding and IQ shows moderation, where there are benefits but they max out after a point; marbling, where formula has pros in addition to cons; or granularity, where breast is best for most babies but not those born underweight. (Don’t worry, I didn’t actually use the words moderation, marbling and granularity during the call.) However, the counsellor was unequivocal – we shouldn’t give any formula. She said that colostrum, the liquid new mothers produce before milk comes in, is enough to line a baby’s stomach. I wanted to feed our son, not just line his stomach, but she was adamant – nature provides everything you need. However, she also mentioned a few times that she wasn't medically qualified, so after hanging up I did a bit more digging.

I went back to the searches I'd done pre-birth. This time, I resolved to read the actual papers, not just go by the Google preview. Seeing the whole picture led to quite different conclusions. For the *Frontiers in Nutrition* study, the full quote was 'On an average, more breast-fed participants had high IQ scores than non-breast-fed participants. These findings agree with ours to some extent. However, because of the small sample size, we could not confirm the significant difference between the breast-fed and bottle-fed groups.'[*] In other words, the difference was so small that it could be due to luck – a sharp contrast to what the Google excerpt suggested.

Even more strikingly, the paragraph containing that quote started by describing a different study which 'concluded that breastfeeding had no significant effect on intelligence (18)'. Footnote 18 referenced a *British Medical Journal* article by Geoff Der, David Batty and Ian Deary.[4] I dug it up; at six pages long, it would be a quick read.

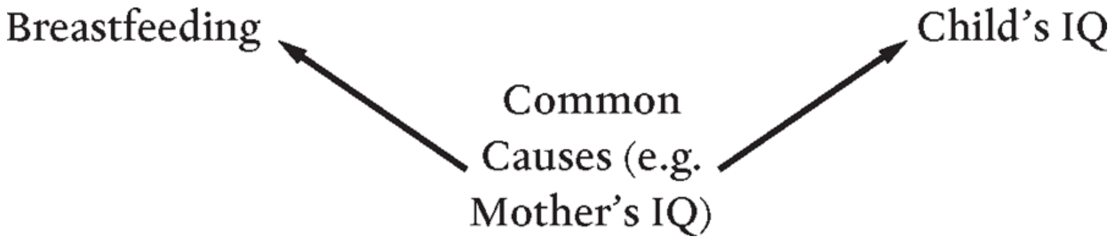
The researchers used the scientific method. They took a random sample of 5,475 children aged between 5 and 14, divided them into test and control groups (breast-fed vs bottle-fed), and compared their IQs. Geoff, David and Ian found that breast-fed kids had IQs 4.69 points higher than their bottle-fed peers – a statistically significant difference. This reinforced the first two studies I'd found on Google, as well as conventional wisdom. All pointed to the following relationship:

Breastfeeding → Child’s IQ

But whether a child is in the test (breast-fed) or control (bottle-fed) group isn’t random, because breast-and bottle-fed babies differ in many other ways. Breastfeeding is difficult for mothers who can’t afford to take time off work or don’t work in an office where they can pump their milk and refrigerate it; even at home, it can be exhausting without family support. Indeed, the authors found that breastfeeding mothers tended to have higher IQ themselves; they were also older and more educated, had a better home environment and were less likely to be poor or to smoke. These factors are *common causes* – they affect both the input (feeding method) and the output (IQ).[†] They, not the feeding method, could be what led to the differences in IQ. Common causes are a rival theory for why breastfeeding and IQ are correlated.

After the authors *controlled for* the mother’s IQ – stripped away the differences in child IQ that can be explained by the mother’s IQ – the gap plummeted from 4.69 to 1.30 points. That’s still statistically significant, but when the researchers also controlled for other common causes such as smoking, family poverty and the home environment, the difference became an insignificant 0.52.

The diagram below illustrates – there’s no link between breastfeeding and child IQ; instead, the common causes affect both.



Of course, we cared about many outcomes other than IQ. But we had to make a quick decision for the next feed. This study had already highlighted the importance of common causes, so we wondered if they might also drive the other outcomes frequently touted. We gave Caspar a bottle of Aptamil Profutura First Infant Milk, which he guzzled down, and then used a combination of breast and bottle while I investigated the other supposed benefits of breastfeeding more carefully.

My search led me to the book *Cribsheet: A Data-driven Guide to Better, More Relaxed Parenting, from Birth to Preschool* by health economist Emily Oster.[‡] *Cribsheet* uses the highest-quality scientific evidence to re-examine widely held beliefs about parenting and see which stand up to scrutiny. The book has a whole chapter devoted to breastfeeding, which starts out by listing all the claimed payoffs: short-term benefits to the baby (such as fewer infections, fewer allergic rashes, lower risk of sudden infant death syndrome), long-term benefits to the baby (less diabetes, lower risk of obesity, higher IQ), and benefits to the mother (lower risk of post-natal depression and osteoporosis, better bonding with your baby).

After poring through all the evidence and focusing on studies that control for common causes, only a few of the outcomes hold up,[§] including none of the ten long-term gains to the baby. Those that remain are still important, but they're not as extensive as often claimed and so the decision is no longer so cut and dried. When balanced against the benefits of the bottle, we decided to combination-feed from then on – breastmilk as the first choice, but formula if Caspar needed a top-up or my wife wanted a break.

Data is not enough

Everyone knows the phrase 'Correlation is not causation', but not necessarily why. The breastfeeding studies show us exactly why. They did everything we've recommended so far. They started with a hypothesis (breastfeeding affects IQ) and gathered a random sample with different levels of the input (feeding method). There was scant risk of data mining, since there's little ambiguity on how to measure IQ or classify whether a baby was breast-fed. But no matter how robust the correlation between breastfeeding and IQ, they didn’t show causation because it might be driven by common causes.

This example shows us that *data is not evidence* : it may not be *conclusive* if it’s consistent with alternative explanations. Data is just a collection of facts. Evidence is data that allows you to distinguish between hypotheses – to support yours and rule out alternatives. Even if a study is grounded on rock-solid data, it may still contain lies.

We’ve seen in previous chapters how getting from data to evidence requires a control group with different levels of the input – CEOs that aren’t adopted, companies that don’t start with *why* and investors who don’t trade frequently. However, different inputs aren’t enough if they’re *endogenous* – if they’re not random. In other words, they’re affected by the same common causes that drive the output. There are three main reasons why an input might be endogenous: it’s a voluntary choice, it’s correlated with other traits or it’s the outcome of another process.

Many studies involving people have a *voluntary choice* as the input. People don’t make choices randomly – mothers don’t suddenly wake up and dump the bottle; instead, their decisions are based on various common causes. Perhaps a more supportive home environment helps mums to breastfeed, and this environment also boosts child IQ. Voluntary choices also plague ‘A new concept in the treatment of obesity’, the paper on which Atkins founded his diet.[11] People don’t just randomly start dieting. They consciously do so because they want to lose weight; this desire might also drive them to exercise, and it’s the exercise, not the diet, that causes the weight loss. The common cause is the resolution to get fit.

Other research on people examines their traits. While traits aren’t voluntary choices, they’re still endogenous, because they’re *correlated with other traits*, and it could be those other traits that are driving the output. A respected newsletter opened with: ‘Leaders who prioritise the emotions of their employees are 2.5 times more likely to be successful than those who don’t, according to a new study by EY and Oxford University’s Saïd Business School.’ There was no link to the study, and neither the EY[5] nor Oxford[6] press releases contained one, so nobody could check how they captured whether a leader prioritizes employees’ emotions. But even if the researchers had a perfect measure, it’s probably correlated with many other characteristics. Perhaps leaders with good EQ also have high IQ and it’s IQ that leads to success. Or, bosses who care about their colleagues’ feelings are better people managers in general and it’s those other aspects of people management – mentorship, coaching and empowerment – that drive achievement.

The third reason an input can be endogenous is if it’s the *outcome of another process* : certain factors cause companies, cities and countries to be the way they are. In 2020, as the coronavirus crisis unfolded, a study claimed that greater air pollution was associated with more COVID cases and deaths, with causal statements such as ‘poor air quality increases the lethality of COVID-19’.[7] Newspaper headlines broke out, such as ‘“Compelling” evidence air pollution worsens coronavirus – study’ with the strapline ‘Exclusive: best analysis to date indicates significant increases in infections, hospital admissions and deaths.’[8]

Cycling is my main form of transport but often unpleasant in central London, so environmentalists like me wanted to believe this result and use it to lobby for anti-pollution action – yet there’s a major flaw. High pollution isn’t something that cities voluntarily choose, nor a trait they’re born with, but it’s still endogenous as it’s an outcome of another process: certain factors cause a city to be polluted, and these same factors could also affect coronavirus. For example, population density is a common cause that both increases pollution and accelerates the spread of COVID. You don’t need statistical wizardry to come up with this alternative explanation, just common sense – but it’s often switched off when confirmation bias is at play.

Why this matters

Why is it so important to distinguish correlation from causation? Taken literally, study results are *descriptions* about the world. Breast-fed babies have better outcomes, people who start diets lose weight and more polluted cities transmit coronavirus faster. These statements aren’t inaccurate – breast-fed kids do have higher IQs – and if you use the findings to simply describe how the land lies, that's fine. Countries beginning with U (such as the United States and United Kingdom) are richer than the world average, but no one would ever use this to suggest that changing your country's name would boost its GDP.

The problems arise when we extrapolate from a description to make *predictions* about the world. If breast-fed babies have higher IQs, then a mother might think that ditching the bottle will set her kid up for success. However, if the higher IQ is driven by common causes such as not smoking, then switching the feeding method won't have any effect. In fact, it may backfire by distracting her from the true solution – giving up cigarettes.

The temptation to extrapolate is so great that even the world's most respected companies make this mistake. In 2017, McKinsey released an influential study which claimed that companies that act more long-term (for example, by investing more) deliver better long-term performance. That's a description of what they found, but they turned it into a bold prediction: if all US companies started acting more long-term, the economy would grow by an extra $3 trillion over the next decade. Given such a blockbuster result, *Harvard Business Review* published an article on the study with a blockbuster title: 'Finally, proof that managing for the long term pays off '.

Several common causes could be at play here. One is the industry. Being in a growing industry, such as electric cars, will cause a company to invest more, and the same industry growth will also boost its share price. Tesla invests more than Imperial Tobacco, and Tesla performs better, but its superior returns are more due to its sector than how much it spends. If Imperial Tobacco tripled its outgoings, its returns would fall rather than rise, since there aren't great investment opportunities in a declining sector like tobacco. Luckily, enough discerning readers complained to *HBR* that they took one step down the Ladder of Misinference, changing the title to 'Finally, evidence that managing for the long term pays off '.[9] But it should have been two steps back, since the study only had data, not evidence.

As readers, we should be wary ourselves of jumping from a description to a prediction. We previously saw how headlines such as 'People who do X are more successful', 'Companies with Y are more profitable' and 'Countries with Z are happier' are often meaningless, as there's no mention of statistical significance – the outperformance could be entirely random. And there's a second problem: even though the statements strictly describe correlations, readers often interpret them as causation – if you do X, Y or Z, you'll also flourish. However, there could be a ton of common causes correlated with that behaviour or trait that are responsible for the outcomes. In all these cases, we should ask ourselves: *Do the input and output share a common cause?*

Headlines like these often have problems beyond just luck and common causes – they suffer from flaws highlighted in every chapter of this book. In April 2023, a

LinkedIn post trumpeted: ‘After 11+ years the most #trustworthy public companies continue to outperform,’ accompanied by a graph showing that trustworthy companies beat the market. The writer tagged me, hoping I’d be impressed, but my reaction was anything but. The post was by the founder of ‘Trust Across America’, so she wanted the result to be true (Chapter 1); it gave the black-and-white impression that trust always pays off (Chapter 2). It didn’t mention how the study measured trust (Chapter 3),[#] there was no test of whether the outperformance was statistically significant (Chapter 4), we didn’t know how many other ways the researchers tried to measure trust (Chapter 5), and trustworthy companies might differ along other dimensions, such as having a great leader (this chapter).[**]

Yet people immediately piled on the platitudes, with comments such as ‘Impressive data’, ‘SOLID stuff ’ – probably without even clicking on the link to the study. And this is unfortunately the rule, not the exception. The practice of taking a set of people, companies or countries and showing that they beat their peers is extremely common – it pervades articles in business newspapers, studies by management consultancies and talks by self-appointed gurus. Whenever we see such claims, we should first pause and ask ourselves whether any of the six problems we’ve encountered so far apply.[††]

Bold predictions make the biggest splashes. Part of the reason why the McKinsey study went viral, despite its flaws, is that it promised that $3 trillion would fall from the sky if companies would only follow its recommendations. Similarly, if an article pledges to increase company profits by 142% or athletic performance by 79%, you'd better read it urgently or you'll lag behind your peers; if a book swears success with just a four-hour workweek, you'll drop your coffee cup to order it.

But attention-grabbing numbers should have the opposite effect – they should prompt us to stop and think whether they're plausible. If a simple technique really did increase profits by 142%, then all those not practising it would quickly go out of business. If you could reach the top by grafting just four hours a week – or even twenty-four – then almost all the world's leading executives, scientists and movie directors would be needlessly overworking. In contrast, a report on how to increase your metaphorical batting average from 0.280 to 0.320 is unlikely to hit the headlines, but there are far more likely to be foundations beneath the house.

Just like Goldilocks wanted her porridge to be neither too hot nor too cold, the most convincing statistics are neither too high nor too low. In Chapter 4, we explained how the link between an input and an output needs to be statistically significant – large enough that it's unlikely to be due to luck. Now we need to add a second condition: it can't be so large that it breaks the boundaries of believability.

We’ve shown that to move from data to evidence we need to control for common causes. But what does ‘controlling for’ actually mean, and how do we do it? We’ll now scrutinize one of my own papers to see how I addressed this pesky problem.

Eliminating the alternative suspects

‘Just one more hour,’ I told myself, as I struggled to keep my eyes open. It was 2 a.m., but I was still in the Morgan Stanley offices. My boss wanted the first draft of a presentation on his desk by 9 a.m. We were pitching to a client the idea of taking over a competitor, and I’d burned the midnight oil calculating every cost and benefit from doing so. The final piece in the jigsaw was a ‘waterfall diagram’ to show just how ingenious our idea was. It began with the client’s current value, added or subtracted each gain or loss, and ended up higher than where it started.

I sketched what I wanted the waterfall to look like and trudged to the lift to take it to the graphics department. I’d then check the rest of the presentation for catastrophic errors, such as different graphs having slightly different line thicknesses, while the designer chugged away. If all went well, I’d be heading home in an hour.

As I reached the lift, I was struck by a poster I hadn’t seen before. It showed a set of stones, each perfectly balanced one on top of another. And it was quite a balancing act – the stones weren’t lying horizontally but vertically, each on its thin end. Even in my slumbering state, I couldn’t help but admire the beauty of the image. I read the text and saw the poster was advertising Morgan Stanley’s new ‘balanceworks’ programme, to encourage its employees to have a better work–life balance. To make the programme particularly snazzy, there were no capital letters in ‘balanceworks’, and the ‘balance’ was in a bolder font than the ‘works’ – a rare example of inconsistent thicknesses being allowed within these walls – to convey how our bosses cared even more about balance than work.

Fine words, but I smiled wryly about seeing such a poster at 2 a.m. Yet Morgan Stanley wasn’t alone – every investment bank, law partnership and management consultancy makes grand claims about people being their greatest asset and being treated like family. Yet few seem to walk the talk, instead acting as if the route to profits is to squeeze as much as possible out of every last worker.

Might they be wrong? If employees aren’t so overworked, they’ll make fewer mistakes; if they’re monitored less, they won’t use their freedom to shirk but will innovate. When I arrived at MIT for my Ph.D., I decided to test this hypothesis – whether companies that treat their workers better also perform better. I measured employee satisfaction using the list of the 100 Best Companies to Work for in America,[‡‡] and firm performance using shareholder returns. This led to the paper that Xinyi approached me about in Chapter 5.

My main headache was common causes. For example, Google often tops the Best Companies list, and Google has enjoyed exceptional returns. However, this may be nothing to do with employee satisfaction, but it being a tech company. Tech firms tend to have happy employees because the work is creative and independent, and the tech industry has performed strongly. In contrast, coal mining has harsh working conditions and is in decline. The common cause is the industry.

Addressing this doesn’t seem too tricky. You create a control group by comparing each Best Company with non-Best Companies in the same industry – Google with other tech firms, Marriott with rival hotels, and so on. I found that the Best Companies as a whole beat their industry peers by 2.3% per year, or 84% over the 28 years.[§§] [10]

But the problem with common causes is the list can be pretty long. What about size? Google might be special not only because it’s a tech company, but also because it’s a giant one. Large firms beat small firms from time to time; for example, they’re more resilient in a downturn. Another common cause is recent performance due to a phenomenon known as ‘momentum’ – stocks that have done well recently tend to continue their winning streak. A third is dividends. Perhaps companies that can afford high dividends can also treat their employees well – and it’s high dividends, not happy workers, that cause the stock price to soar. If I had to compare a Best Company to a non-Best Company in the same industry, with similar size, recent performance, dividends and several other common causes not listed here, there might be no companies that match on every dimension. I’d have no control group.

The solution is to use the *regression* (best-fit line) we encountered in Chapter 5. Back then, we only had one input, and we saw how regressions allow this input to be anything you like – the number of female directors can be 0, 1, 2, 3, 4 or 34, rather than just high or low. Here we’re interested in a related advantage of regressions: you can have as many inputs as you like – you can link stock returns to Best Company status, industry, size, recent performance and dividends all at the same time.[||||] Those extra inputs are known as *controls*.[##] The terminology is deliberate – you ‘control for’ a factor by adding it as a control to your regression.

Now, the regression tells you how much the output rises if you increase a single input *without changing any of the other inputs* – for example, if a firm switches from being a non-Best Company to a Best Company, while keeping fixed its industry, size, and so on. This method allowed me to find the link between being a Best Company and returns even without a perfect control group – without a non-Best Company that’s exactly identical along every other characteristic.[***]

A regression is the best way to control for common causes. It’s simple – schoolkids study it in A-level Maths – and the results are easy to interpret. But despite this simplicity, some influential studies and books opt for invalid solutions. Let’s take a look at one of them.

Out of control

I started at St Paul’s, my secondary school in London, at age thirteen, like nearly all my classmates. A couple of boys transferred in at sixteen; one, whom I’ll call Max, joined my A-level Maths class and was insanely smart. While nearly every football fan in the school supported a London team like Arsenal or Spurs, or chose the glory of Manchester United or Liverpool, Max followed his local club, Crewe Alexandra. I had a season ticket at Reading, since I lived there; you’d never voluntarily support them otherwise. Like Crewe, Reading languished in the division below the Premiership, so we had a natural connection. Chats in breaks and lunches didn’t often venture into politics, but Max and I had similar views on inequality and might have been the only boys in our year to celebrate the Labour Party winning the 1997 election.

We kept in touch despite going to different universities but lost contact when I went to the US for my Ph.D. Soon after returning to the UK, I became involved in policy discussions on executive pay. I shared some of my articles on social media, and Max took a very critical position on high CEO salaries. He referred to a book titled *The Spirit Level* to back up his arguments, but I said I hadn’t read it. A few days later, I received a copy in the post, thanks to

Max. It bore the subtitle ‘Why Equality is Better for Everyone’.

Everyone? Even the wealthy? That was a bold claim, and I wondered if it was preying on black-and-white thinking. But the front cover had an endorsement from the *Economist* – ‘The evidence is hard to dispute’ – and there was similar praise from Ed Miliband (the then head of the Labour party) and David Cameron (the then Conservative Prime Minister). That leaders on both sides endorsed a book on a politically charged topic like inequality must mean that the evidence was indeed indisputable.

The authors, Richard Wilkinson and Kate Pickett, gathered data on a country’s physical health, mental health, obesity and eight other outcomes. They drew eleven best-fit lines, each relating income inequality to a different output, and claimed that inequality worsened every single one. They argued that inequality *caused* the negative consequences, as indicated by the ‘Why Equality is Better for Everyone’ subtitle.

But there could be a whole host of common causes. An obvious one is poverty – poorer societies tend to be more unequal, and poorer societies also have worse health (and worse other outcomes). If it’s poverty, not inequality, that causes ill health, governments should focus on improving the incomes of the poor; there’s no benefit to clipping the wings of the rich.

Wilkinson and Pickett claimed to address this concern by plotting another best-fit line linking poverty to health and finding that its slope was insignificant. But any schoolkid studying A-level Maths could tell you this isn't a valid solution. This new graph showed that *poverty* is *unrelated* to health – but that's completely different from the result the authors were parading. Their grand claim was that *inequality* is *related* to health. To make this claim, they needed to show that it remains true *while controlling for poverty* – while holding poverty constant.

The only way to do this is by conducting a single regression of health on both poverty and inequality together, not separate regressions. This compares the health of two countries with the *same* level of poverty but *different* inequality – it holds poverty constant and changes inequality in isolation. Indeed, a study by Simone Rambotti does exactly this and finds that, when you control for poverty, the link between inequality and health becomes much weaker.[11]

Max was now a university lecturer in statistics – yet confirmation bias caused him to forget his schoolboy statistics when reading *The Spirit Level*. And it wasn't just Max, David Cameron and Ed Miliband. Many other readers were deceived, since, like me, they may dislike inequality. The next person to bring up this book, in a discussion with me on CEO pay, was the then Executive Director of the Royal Statistical Society, who you'd hope should understand statistics yet thought the results were flawless.

The person on the street also fell for it. The Amazon fivestar review with the most likes, by 'penpushing1', is entitled ‘Timely, devastating confirmation of what we all, at bottom, knew anyway’. And that’s the problem – ‘penpushing1’ gave it full marks because it backed up what he/she thought to be true. The negative Amazon reviews mentioned a book that claimed to debunk it, *The Spirit Level Delusion* by Christopher Snowdon, so I looked at that book’s page also. The one-star review with the most likes, by ‘Griffo’ ended with ‘What you think of this book will largely depend upon which side you are on. No prizes for guessing where I stand.’ No prizes for guessing which bias this demonstrates.

Looking under the lamp post

I’ve played up regressions by highlighting how they can include as many controls as you like – but you can only control for what you can observe. For breastfeeding, you can measure the mother’s IQ, but what about patience? More patient mothers might be more likely to breast-feed and also to read to their babies, boosting their IQ. You can’t control for patience because you can’t measure it. Since regressions can only include what you can see, they’re like the proverbial drunk who looks for his keys under a lamp post because that’s where the light is, rather than in the bushes where he’s dropped them. Where the common causes are unobservable, we’ll need to use other tools to move from data to evidence, which we’ll discuss in Chapter 7.

But there's a silver lining. You can't control for every common cause, but you don't have to. Something is only a problem if it's correlated with *both* the input and the output – it has to be common to both. If it only affects the output, omitting it from your regression doesn't affect the slope.[†††] This is important, since you'll never be able to control for every single factor that affects the output. The health of a country might depend on its national diet or the quality of its fitness facilities. It's very difficult to measure these – is having pad thai as your national dish better or worse for your health than it being paella? – but it doesn't matter if they're not linked to inequality.

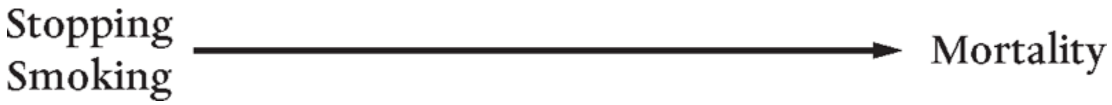
Sometimes a factor might be correlated with the input, but in the 'wrong' direction and so work against your theory. Extreme weather probably worsens health; it could also reduce inequality as a weather-related disaster makes everyone poorer. You can't dismiss Wilkinson and Pickett's negative correlation between inequality and health by saying they didn't control for extreme weather. Meteorological calamities would lower both inequality and health, moving them in the same direction and causing them to be *positively* linked – the opposite of what Wilkinson and Pickett found. In contrast, poverty increases inequality but decreases health, leading them to be negatively correlated, which is what *The Spirit Level* documented. Thus, poverty is an alternative explanation for Wilkinson and Pickett's result, but weather is not.

You only need to worry about a common cause if it's a rival theory. If a study fails to control for a factor but it's likely to be either uncorrelated with the input or correlated in the wrong direction, it's not a problem. This again highlights the importance of common sense – asking how a missing factor might be linked to the input in real life. If we don't like a study, we might engage in blinkered scepticism and complain 'The authors didn't control for the CEO's hair colour' – but they never needed to.

We've so far discussed the problem of common causes, how to deal with them, how not to deal with them, and whether you even need to deal with them. There's a second, and final, reason why correlation may not be causation, which we'll now come to.

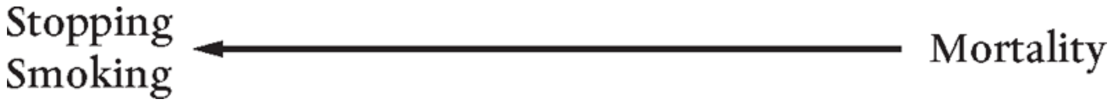
When the tail wags the dog

Nicotine patches, vaping, gum, hypnosis, acupuncture . . . even taking up knitting. There are as many routes to stopping smoking as there are brands of cigarette. But none of them is a silver bullet; smokers often have to try several and experience numerous false dawns before finally kicking the habit. Even if they succeed, the evidence isn't encouraging. Some studies find that stopping smoking leads to a *greater* likelihood of dying.[12]



What’s going on here? Hopefully by now you’re on your guard that correlation needn’t be causation – it’s unlikely that giving up cigarettes drove the higher deaths. However, it’s not clear what the common causes might be. Most things that encourage quitting should *reduce* mortality: they’re correlated in the wrong direction so we don’t need to worry about them. A New Year’s resolution to live a healthier lifestyle would lead someone not only to stop smoking but also to eat more healthily and exercise.

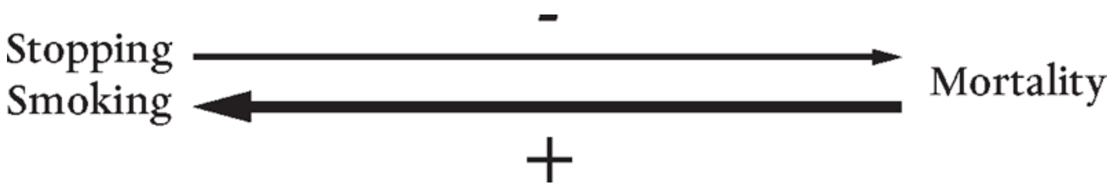
Here, the problem isn’t that the input is driven by common causes – instead, it’s driven by the output itself. This is known as *reverse causation*, and illustrated below:



When a doctor tells smokers they’re at high risk of lung cancer, many bite the bullet and quit. The fear of death causes someone to stop smoking, rather than stopping smoking leading to death. Similarly, patients often become sicker after going to the doctor. That’s not because going to the doctor makes you unwell; instead, you visit her when you have an illness oncoming.

Both cases are examples of the *post hoc ergo propter hoc* fallacy ('after this, therefore because of this'). If the output followed the input – you get sick after you visit the doctor – we think that the latter caused the former. Instead, you often choose the input (going to the doctor) in anticipation of a particular output (you feel the first symptoms of sickness). Opening an umbrella doesn't cause it to rain.

In the above cases, reverse causation changes the sign of the correlation, from positive to negative, or vice versa. Quitting smoking does lower mortality, but we don't see this in the data because mortality risk increases the likelihood of quitting by such a large amount. The latter drowns out the former, leading to a positive correlation overall.



In other cases, reverse causation strengthens the correlation rather than causing it to flip, so it makes you think there's an effect when there isn't. In *The Spirit Level*, inequality might have no impact on health; instead, ill health causes inequality. If people are sick, they can't go to work and have to spend their savings on medical bills.

Reverse causation is a particular problem if the input is self-reported. Cast our minds back to the Ericsson paper, where students estimated how much they practised the violin up to eighteen years ago. Since their memories were blurry, the violinists in the top group may have thought, ‘I’m good, so I must have worked really hard to get here,’ and reported a high number of hours. The mediocre ones would have been reluctant to claim they practised a lot, otherwise they’d have to admit they slogged away to no avail. How good you are now affects how much you recall you rehearsed – it’s success that leads to reported practice, rather than practice leading to success. That’s similar to Barry Staw’s study, where groups who thought their forecasts were accurate claimed their team dynamics were cohesive.

In all these cases, the problem is not difficult to spot – you ask: *Might the output affect the input?* If so, and if the direction of the effect is the same as the result that’s being paraded, then reverse causation is at play and data is not evidence.

In a nutshell

- *Data is not evidence* because it may not be *conclusive.* Correlation may not be causation due to *common causes* that drive both the input and the output.
    - A mother’s IQ may increase both the likelihood of breastfeeding and child IQ, rather than breastfeeding improving child IQ.

    - Statements such as ‘People who do X are more successful’ are meaningless, because people who do X may differ in many other ways.

- There are three main reasons why an input might be *endogenous* (non-random) – why it might be affected by the same factors that drive the output:

    - It’s a voluntary choice (following a diet).

    - It’s correlated with other traits (bosses with high EQ also have high IQ).

    - It’s the outcome of another process (air pollution is caused by population density).

- Ask: Could something else have caused the output? Might this ‘something else’ also be correlated with the input?

- If the input is endogenous, a correlation remains an accurate *description* of the data but it can’t be used to make a *prediction*. We should be particularly wary of predictions that promise implausibly large effects from taking a certain action.

- To address common causes, control for them in the same regression.

    - You can’t control for common causes that are unobservable. This doesn’t matter if they would work against your results, i.e. drive the input and output in different directions to what you find.

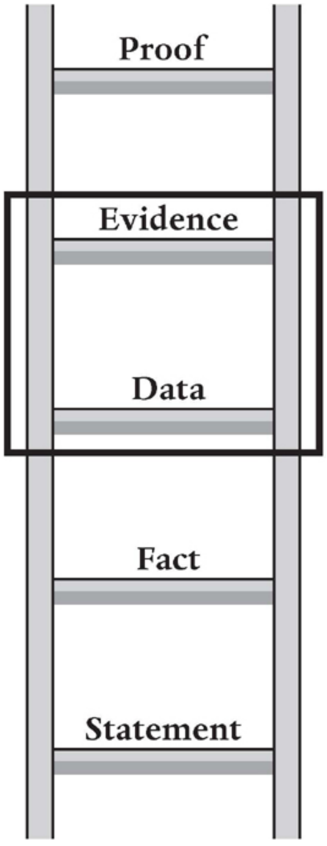
- A second reason why correlation is not causation is *reverse causation* : the output affects the input.

Part II so far, and this chapter in particular, may have presented a bleak picture. It seems we can’t show anything. Even if we first form a plausible hypothesis, then test it using a representative sample controlling for all observable common causes, and finally demonstrate robustness to different measures and sampling criteria, we still have the problem of unobservable common causes or reverse causation.

But having gone through the Pandora’s box of all the ways we might be deceived by statements, facts and data, there’s one thing still left in the jar – hope. The next chapter explains how you *can* turn data into evidence – rule out alternative explanations and support a hypothesis rather than just showing consistency with it.

7

When Data is Evidence



After you’ve resolved the breast versus bottle conundrum, cheered your child’s first words and beamed at his baby steps, another big parental decision is schooling. Just as breast-and bottle-feeding have their staunch supporters and ardent adversaries, there are equally strong views on how much choice to give parents in school selection.

Some argue that competition between schools encourages good performance, just like the free market spurs companies to offer the best products. This benefit of choice is the rationale behind charter schools in the US and academies in the UK, which are run independently rather than by a local authority.

Opponents are just as vocal. In the free market, people make decisions individually, but education takes place collectively – kids learn from each other. Under free choice, parents with smarter kids might send them all to the same school, preventing cross-ability learning. Or, birds of a maths and science feather flock together, meaning they don't develop in humanities and arts. What's individually optimal for each child may be collectively undesirable for society.

Who's right? As with any debate, we look at the evidence. In most countries, it's much easier to get your kids into a school in the district where you live, so parents reside in the districts with the best schools. In some cities, this isn't too difficult. In Boston, there are seventy school districts within a thirty-minute commute of the downtown area.[*] Parents can take their pick – they can live in any one of those seventy and still have an easy trip to the office. In contrast, the Dade County school district in Miami covers virtually all the metropolitan area, so you're stuck with it unless you can stomach a long commute. The number of school districts within a metropolitan area is therefore a good measure of competition, and hence parental choice.[†]

Imagine we gather data from hundreds of school districts and run a regression. We find that child performance is higher in cities with greater school choice, for example Boston compared to Miami. If we're a free marketer, we're tempted to accept this correlation as causation – as evidence that competition causes higher performance.

This is when our rational System 2 should kick in and remind us that there are two alternative explanations for any correlation. One is *reverse causation*. Poor child performance might reduce the number of districts: when districts are doing badly, they merge to cut costs. Another is *common causes* : if parents care about education, they'll demand school choice (increasing the number of districts) and tutor their kids at home (improving child performance). You can't measure parents' concern for education, so you can't control for it.

So what do we do? To understand how to move from correlation to causation, let's first see what researchers did to solve an even more important question than our children's education – one of life and death.

How randomness achieves precision

The Age of Discovery saw legendary captains such as Ferdinand Magellan, Vasco da Gama and Sir Francis Drake venture far beyond the maps of the day. Those were the times when European explorers first crossed the Atlantic, Pacific and Indian oceans, conquering and colonizing new lands. They returned to Europe triumphant bearers of wealth and knowledge – but also with heavily depleted crews. For as much as seafaring was glorious, it was also dangerous, with thousands of lives lost to wars, pirates and the weather. Yet none of these came close to the biggest killer of all: scurvy.

Scurvy is a cruel disease. The earliest symptom is a lethargy so debilitating that even the slightest movement is a Herculean task. As the affliction advances, your gums bleed and smell like rotting flesh, your teeth loosen and ulcers break out across your limbs which soon turn to gangrene. Agonizing pain sears through your muscles, joints and bones. When death comes, likely from a haemorrhage in your heart or brain, it’s a welcome mercy.

Two million sailors perished from scurvy between 1500 and 1800 – roughly the period between Christopher Columbus setting sail for America and Richard Trevithick inventing the steam train. The disease was so serious that captains assumed it would claim half their crew on any major voyage. The search for a cure became ‘a vital factor determining the destiny of nations’, according to historian Stephen Bown.[1] Any country that could prevent scurvy would be guaranteed a huge military advantage.

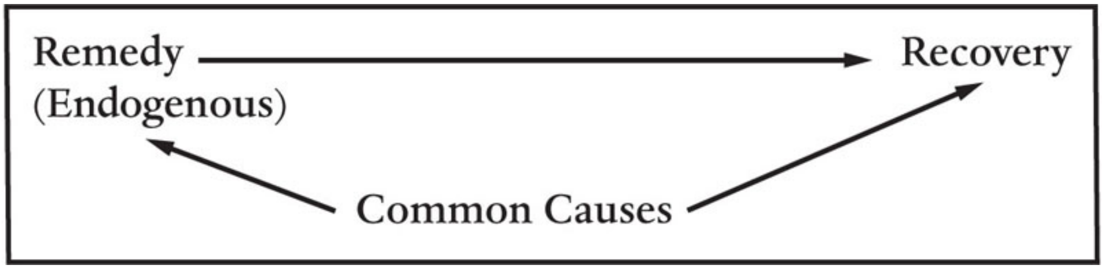
Driven by desperation, explorers tried almost any remedy, ranging from the repulsive to the bizarre. On a voyage to the East Indies, Vasco da Gama ordered his sailors to wash their mouths with their own urine. Other efforts included elixir of vitriol, which was more vitriol than elixir, as the main ingredient was sulphuric acid, and an ‘electuary’ – a curious concoction of garlic, mustard seed, myrrh and balsam of Peru (sap from the *Myroxylon balsamum* tree).

There was no method to this madness. Clutching at straws in desperation to find a cure, captains had no headspace to create a system to figure out what worked. Even if, say, the electuary led to a greater recovery rate than the elixir, there’d have been many common causes. Perhaps those given the latter were suffering from the most serious cases of scurvy – you’d have to be desperate to start swallowing sulphuric acid – and so they’d have died anyway. Alternatively, it might have been the explorers who made the decisions. If captains who prescribed the electuary had the most success, it could be because those who could get their hands on scarce ingredients were able to take care of their sailors in other ways.

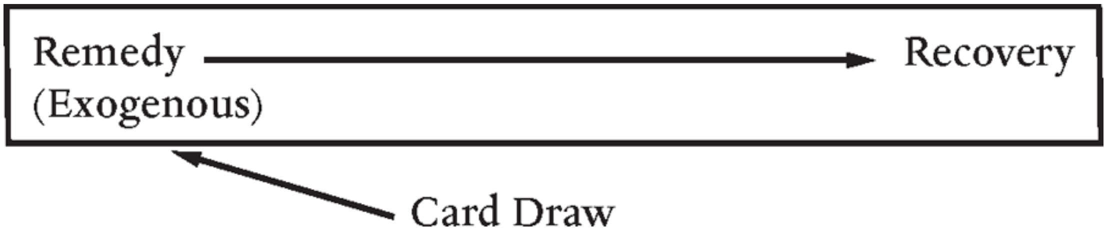
All these alternative explanations arise because the remedy is *endogenous* – it’s either chosen by the sailors or prescribed by the captains. This choice could be correlated with several common causes, such as the severity of the disease and the resources of the explorer.

In 1747, James Lind, the ship's doctor on the HMS *Salisbury*, addressed this problem by making the remedy *exogenous* – by making it random, and so unrelated to any common cause. He arbitrarily assigned the therapies so that neither the sailors nor the captains had any choice in who got what. Lind divided twelve scurvy patients, whose 'cases were as similar as I could have them', into six pairs. Each pair was given a different remedy: twenty-five drops of elixir of vitriol three times a day, the electuary, one quart of cider, two spoonfuls of vinegar thrice daily, half a pint of seawater, and two oranges and one lemon.

The diagram below illustrates this. Before Lind's breakthrough, the remedy was endogenous. It's inside ('endo') the system of relationships between the common causes and the output, and so it's driven by the same factors that affect the likelihood of recovery.



We don't know how Lind did his random assignment, but let's say he drew from a deck of cards. This makes the remedy exogenous because it's now steered by something outside ('exo') the system – whether you get the jack of clubs or the two of diamonds has nothing to do with the sailor's chance of recuperation. That's why the second diagram has no direct link from the card draw to recovery. Any link between the remedy and recovery can't be because the card draw affects both – instead, it must be because the remedy causes the recovery.



The pair who were randomly assigned the citrus fruits recovered so quickly that one returned to active duty and the second was able to look after the other patients. Lind's finding was the first evidence that citrus fruits cured scurvy, a discovery that ended up saving millions of lives.

More broadly, it's the first documented example of a *randomized control trial* (RCT). The studies we've encountered in Part II so far have been *observational studies*, where you observe what people or companies do, based on existing data, and try to make inferences – but you're always plagued by common causes. The solution is an *intervention study*. Rather than seeing what people do, you tell them what to do, randomizing who gets what. If the input is arbitrarily assigned, nothing's causing it, so there can't be any common cause.[‡] Then, correlation does imply causation, and so *data is evidence*.

Why something is better than nothing

RCTs are the gold standard in showing causation, and the methodology has evolved further since Lind's time. The one weakness in Lind's study is that even causation might not be enough to move from data to evidence. Even if citrus fruits cause a recovery, there could still be alternative explanations as to why. It might not be due to their nutritional content but a placebo effect. Those given the oranges and lemons might have believed they'd be cured, because citrus fruits seemed the most plausible of all the remedies, and it was this psychological effect that made them better. In contrast, those prescribed sulphuric acid thought they'd drawn the short straw and were doomed to death.

Austin Flint, a doctor who later became President of the American Medical Association, found a clever way to deal with this problem. One of his studies suggested that a conventional drug cures rheumatism – those prescribed it recovered faster than a control group given nothing. Flint was concerned that the improvement was due to psychology – patients who received a treatment believed they'd get better. Then, the control group wasn't a fair control group, as they got nothing at all and so didn't expect to recover. What he needed was a control group that was given *something*, but just with no medical properties.

In 1863, Flint conducted another experiment. He gave a heavily diluted extract from the quassia plant, which has no medicinal value, to thirteen patients. They improved just as much as in Flint's earlier study, leading him to conclude that the conventional drug had no effect. This is the first known use of a placebo, where the control group is given a dummy treatment – so they don't know they're the control.

While Flint tested the treatment and placebo in separate trials, his innovation paved the way for the 'blind' RCTs that are used – indeed, required – in clinical trials today. Researchers recruit volunteers and randomly give half of them (the *test group*) a drug, and the other half (the *control group*) a placebo. Subjects are blind to what they're given, so any outcomes can be attributed to the medical effects of the drug rather than the psychological effects of simply taking something.

The success of RCTs in medicine has led to them being used in other fields. One influential study ran an RCT to investigate racial discrimination. In the US, African Americans are twice as likely to be unemployed as Caucasians; even if they have a job, they earn 20% less.[2] One interpretation is discrimination, but defenders of the status quo argue that African Americans may be less qualified.[§] The common cause is ability – perhaps employers make job offers purely based on ability, but because ability is correlated with race, this leads to them hiring more Caucasians.

Economists Marianne Bertrand and Sendhil Mullainathan pitted these rival theories against each other with an innovative RCT.[3] They created a résumé bank by taking CVs posted on two recruitment websites by people looking for four types of jobs[‖] in Boston and Chicago. Marianne and Sendhil classified each résumé as high or low quality based on the applicant’s education and work experience.[#] They then scanned all job adverts in the *Boston Globe* and the *Chicago Tribune* across the four categories. For each ad, they drew four CVs from the résumé bank that fitted the job description, two high quality and two low quality. The key step was to randomly assign a Caucasian-sounding name, such as Emily Walsh or Greg Baker, to one of the high-quality CVs, and an African American-sounding name like Lakisha Washington or Jamal Jones to the other. They did the same with the low-quality CVs, and sent all four off. Overall, they posted 5,000 résumés in response to 1,300 job adverts.

The researchers found that the applicants with Caucasian names needed to send 10 résumés to get one callback but those with African American names required 15 – a statistically significant gap of 50%. To put this into context, changing from an African American to a Caucasian name yields as many additional callbacks as eight extra years of work experience.

Because Marianne and Sendhil used a randomized control trial, their study was powerful evidence of discrimination. The fewer callbacks to African Americans couldn’t be explained by differences in ability, because the researchers held ability constant and changed only the name. This randomization allowed them to demonstrate not an innocent correlation, but a harrowing causation.

A shock to the system

Given the success of RCTs across a variety of fields, you might hope to use them for the debate on whether parental choice improves school performance. You’d randomly merge some school districts (the *test group*), keep your nose out of others (the *control group*) and compare their outcomes.

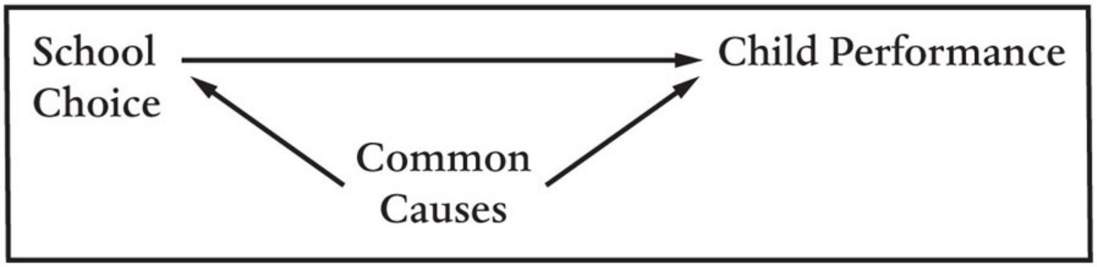
But such an experiment would be both costly and risky. The expense of merging school districts is huge, and if competition does improve performance, thousands of kids in the merged districts will suffer worse education. This is the key limitation of RCTs: they’re effective where you can use them, but sometimes the consequences of putting people in the wrong group are so severe that you can’t take the risk. To test if smoking causes cancer, you can’t recruit volunteers and force half of them to smoke; to verify Belle Gibson’s advice, you can’t sign up cancer patients and order 50% to forgo chemotherapy for clean eating.

In situations like this, you can't conduct an intervention study – you can't intervene and change the input yourself – so what you need is something that already does so in the real world. This is known as an *instrument.*[**] An instrument causes the input to change, but for random reasons that have nothing to do with the output. In simple terms, it's a shock to the system. You can then observe what happens after the input was shocked and conduct an observational study that requires no intervention.

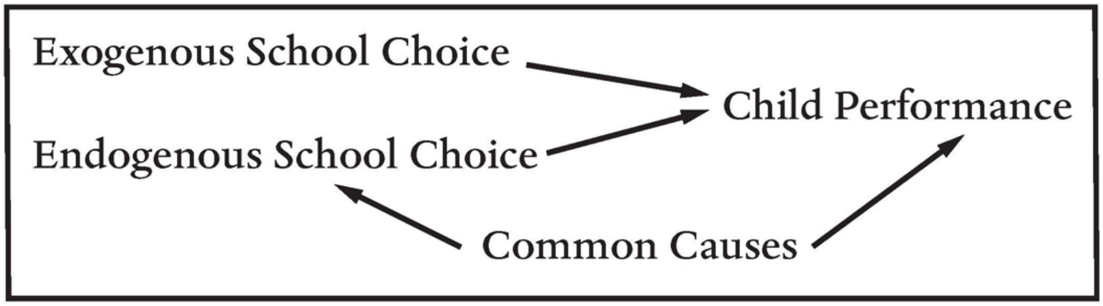
A famous paper by economist Caroline Hoxby used rivers as an instrument for school choice to solve our education conundrum.[4] In the US, school districts were formed in the eighteenth century, when crossing a river was difficult because there were no cars and few bridges. As a result, districts rarely crossed rivers, so that children wouldn't need to do so to get to school. Metropolitan areas with several rivers thus had multiple districts; since districts haven't changed much over time, these areas still have many today.

Hoxby decomposed the input – school choice – into two parts: the exogenous part that can be attributed to the instrument (rivers) and the endogenous part that can't (and instead arises from common causes like parental engagement).[††] Then, she linked only the exogenous part to the output: student performance. Common causes don't affect exogenous school choice – there's no arrow between them – so they can't be the reason for any correlation between exogenous school choice and performance.

Before



After

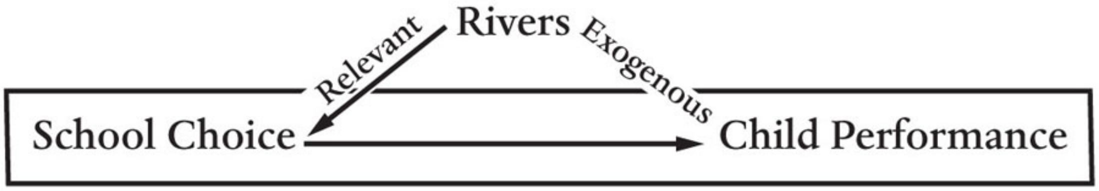


Hoxby pored through data on 30,901 schools across 316 metropolitan areas. She found that metropolitan areas like Boston which naturally had more school districts because they contained more rivers had better schooling outcomes than others like Miami – both short-term ones such as eighth-grade reading score and tenth-grade maths score, and longer-term ones such as the highest level of education achieved and income at age thirty-two. Hoxby’s study is considered a seminal work in the economics of education and has had a significant influence on policy around the world, because the instrument allowed her to demonstrate causation, not just correlation.

Blunt instruments

Just as the ‘superfood’ craze encourages companies to peddle their goods as superfoods, unscrupulous researchers often dress up their work with the magic word ‘instrument’ to claim causation when their methodology is actually bogus. So how do we know whether the emperor has any clothes?

A valid instrument needs to satisfy two criteria. First, it must be *relevant* – it needs to affect the input. Rivers are relevant, as they force a metropolitan area to have more districts and so increase school choice. Second, an instrument must be *exogenous* – it has no effect on the output *except* through the input. Rivers are unlikely to affect a child’s performance other than by changing the number of districts and thus school choice. They don’t directly cause kids to be smarter, nor are there any indirect effects through rivers making parents more engaged. In the diagram below, there’s no arrow from rivers to child performance, so rivers are not a common cause. Any link between school choice and child performance can’t be because rivers cause both.



It’s easy to find instruments that satisfy the first requirement, but not the second. For example, the number of letters written to the local newspaper demanding more school choice is relevant, as pressure from citizens may influence policymakers to increase choice. But letters don’t appear randomly – they’re *endogenous* as they’re a voluntary decision and so they may be driven by common causes. Actively involved parents might write letters, and actively involved parents may also directly improve child performance. The key question to ask is: *Might the instrument be correlated with the output?* Here, the answer is yes: letter-writing is probably linked to kids’ test scores, because eager mums and dads are behind both.

The second requirement means that instruments should have a certain degree of ridiculousness to them. A good instrument should sound crazy – it should appear irrelevant to explaining the output – but this irrelevance is exactly what’s needed for there to be no arrow, making it exogenous. If you’re interested in what determines child performance, it seems wacky to look at rivers. Newspaper letters aren’t so absurd, because they’re written by parents who might also tutor their kids. The crazy instrument is valid; the sensible one isn’t.

Let’s look at a successful instrument in a very different setting to practise how to check if it’s valid. When Rupert Murdoch appointed his son Lachlan Deputy Chief Operating Officer of News Corporation in 2000, critics screamed nepotism. Lachlan became co-chairman in 2014, and now has sole control since his father has finally stepped down.

The naysayers had a point. What if someone outside the family would be a more effective leader? While it wasn’t predestined that Lachlan would take over, it was long known that Murdoch would hand over to one of his children. By restricting the choice set to his six kids, Murdoch overlooked hundreds of potentially better candidates. But like almost any issue, it’s not black and white. Perhaps Murdoch’s children understand the company culture better than any outsider, or they have a longerterm perspective as they want to preserve the family name and reputation rather than extracting short-term profits.

This succession debate matters not only to News Corporation but all family firms. Some of the most influential companies in the world are family businesses, such as Walmart, Ford, BMW, Comcast and Dell. In many countries, particularly in Asia and Latin America, they’re the most common type of enterprise, so the question of who to put in charge is of national importance.

What’s the answer? To find out, we could run a regression linking firm performance to whether the CEO is internal or external. However, even if we found a clear correlation – say, that insider CEOs perform worse – there will always be common causes. If a company is suffering from low employee morale, innovation setbacks or a decline in its brand, then it might have difficulty recruiting outside CEOs and its only option is to keep it within the family. It's these troubles, not the family CEO, that caused the worse performance. Since many of these headwinds are difficult to measure, there's no way to control for them.

Morten Bennedsen and co-authors came up with an inspired instrument to address this problem – the gender of the departing CEO's first-born child.[5] It's relevant, because when the first-born child is male, the departing CEO is more likely to choose a child as his successor. Given gender discrimination, the researchers found that he's more inclined to trust his business to a son than a daughter, and given primogeniture, if he hands over to an heir, it's more likely to be to his first-born. It's also exogenous, because which sperm happens to fertilize the egg has nothing to do with company performance. Remember that valid instruments should be crazy – it seems ridiculous to study the gender of the first-born kid in a paper on firm success – but that's what makes it exogenous. The authors found that companies that are more likely to have a family CEO, because the departing CEO had a male first child, earn significantly lower profits. This suggests that family CEOs *cause* worse performance.

Could the number of children be an alternative instrument? It’s relevant, because families with more children are more likely to keep the company within the family. Is it exogenous? To determine this, recall that we should ask ourselves: Might the instrument be correlated with the output? The answer is yes, and so the instrument is *not* exogenous. The number of children might directly affect performance, because more children means more people to help with the family firm – even though only one can be the CEO, others can help out in different roles. Precisely because the number of children isn’t so ridiculous to look at when studying firm performance, it’s not a valid instrument.

What if we can’t unearth an instrument that satisfies both conditions? If we’re lucky, we might not need an instrument at all. We don’t need an instrument to shock the system if a shock occurs naturally. Let’s look at one of these cases.

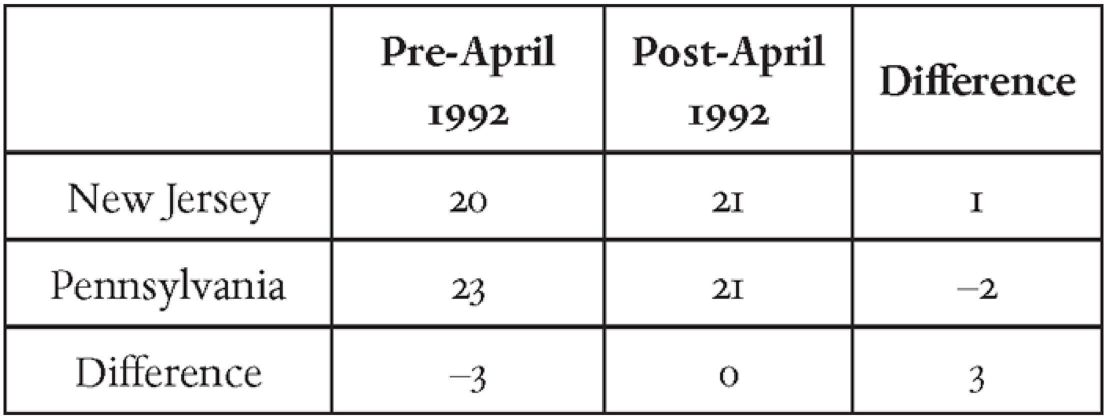
Natural experiments

The school-choice debate pits free marketers against competition’s critics. An even fiercer battleground on which they fight is minimum-wage laws. Free marketers argue that they don’t just harm companies by increasing costs; they also hurt workers. Some employees might be perfectly happy to take a job at below the minimum wage but can’t; higher costs also encourage companies to outsource their work or replace employees with machines. Defenders of the minimum wage claim that what matters isn't the number of jobs, but the number of wellpaid jobs. And even focusing on the number of jobs alone, the effect is unclear – economic theory shows that, in some cases, a minimum wage could *increase* employment.[‡‡]

David Card, who co-won the 2021 Nobel Prize in Economics, and Alan Krueger, who'd have likely shared it had he still been alive, studied the effect of New Jersey increasing its minimum wage from $4.25 to $5.05 per hour on 1 April 1992. They investigated the fast-food industry because most employees in it receive the minimum wage; unlike restaurant and bar staff, they don't earn tips, making it easy to estimate their income. What they found was striking – the higher wage led to *greater* employment. The average number of employees in New Jersey restaurants was 20 before the wage hike and 21 afterwards. While the difference isn't statistically significant, it contradicts the common concern that minimum wages definitely reduce jobs.

How would a staunch free marketer hit back? By using motivated reasoning. Recall from Chapter 1 how Lord, Ross and Lepper found that, if the murder rate didn't fall after capital punishment was introduced, a death-penalty supporter argued that it would have risen even faster otherwise. Similarly, a minimum wage opponent could claim the *counterfactual* isn't 20. Had the minimum wage not been increased, they'd argue that employment would have risen to, perhaps, 22 because the economy boomed. Compared to this benchmark, the actual number of 21 suggests that the wage increase lowered employment.

Rather than speculating what the counterfactual might be, David and Alan estimated it. They studied how fast-food employment changed in eastern Pennsylvania, which is just across the state border and likely affected by similar economic conditions, but which experienced no law change. Here’s the full picture:[§§]



	Pre-April 1992	Post-April 1992	Difference
New Jersey	20	21	1
Pennsylvania	23	21	–2
Difference	–3	0	3

Average employment *decreased* in Pennsylvania from 23 to 21 – a fall of 2. So as a rough ballpark, it’s reasonable to assume that, without the wage rise, New Jersey jobs would have also declined by 2. That they rose by 1 means that employment grew by 3 more than it would have otherwise – a statistically significant *increase*.

This is known as a *difference-in-differences* calculation. The difference in New Jersey’s employment was +1; in Pennsylvania it was –2. The difference-in-differences (the rate at which jobs in New Jersey rose faster than in Pennsylvania) is 1 – (–2) = 3.

David and Alan’s methods might ring a bell. New Jersey fast-food stores are like the test group in a randomized control trial (that experienced a wage increase); their Pennsylvania peers are the control group (that did not). Indeed, the researchers did a second analysis that even more closely reflects RCTs. They focused just on New Jersey stores and selected those paying $4.25/hour before April 1992 – the ones affected by the wage hike – as their test group. The control group was restaurants already offering $5 or more, so the new law made no difference to them; it was as if they swallowed a placebo. The results were similar to the Pennsylvania comparison – employment surprisingly increased among the stores paying $4.25 and decreased among those already giving over $5.

This method is called a *natural experiment*. This is where, in real life, an event happens that randomly divides the sample into test and control groups. Researchers don’t need to intervene and run an actual experiment, because the separation occurs naturally. They can simply observe how the data played out and conduct an observational study. In this case, lawmakers in New Jersey just happened to increase the minimum wage, while there was no change in Pennsylvania. Or, focusing on New Jersey alone, the law change just happened to affect stores paying $4.25, and not those paying $5, because of how it was designed.

Natural experiments are a second way to move data to evidence. You don’t need to find an instrument to shock the input – and worry about whether it’s both relevant and exogenous – if nature’s done the trick for you.[||||]

Unnatural experiments

Just as instruments can be blunt, natural experiments may actually be unnatural. So how do we spot the fakes?

For a natural experiment to be valid, we need the input to be random, just like in an RCT – you can’t choose whether you’re in the test or control groups. The clue is in the word ‘natural’ – it has to be something that nature (or anything else outside your influence, such as a law) decides, rather than something artificial that you manufacture yourself. The key question to ask is: *Can you affect what group you’re in?*

In the McKinsey study linking long-term behaviour to long-term performance, the authors try to claim causation by doing a separate analysis focusing only on companies ‘that experienced the “natural experiment” of changing their outlook during the sample period’. But this isn’t a natural experiment at all. Companies decide whether to increase investment (in which case they end up in the test group) or not (and end up in the control group). As a result, there might be reverse causation – firms increase investment when they have a rosy outlook, so it’s confidence in their future profitability that drives current spending, rather than current spending driving future profitability. There could also be common causes – good managers have great ideas, so they invest more, and good managers also improve performance. Yet the magic words ‘natural experiment’ deceived readers into thinking that the authors had found causation.

The power of common sense

Instruments and natural experiments are powerful when they’re valid, but unearthing genuine ones is very difficult and many fakes abound. So what’s Plan B – is there another way to get from data to evidence if we can’t find a silver bullet? The alternative is to use plain old common sense. While it’s less conclusive than the above methods, it still helps us move part of the way towards evidence, even if it doesn’t take us right to the door. You can use common sense to help your friend or hurt your enemy – run common-sense tests that either support your theory or rebut rival theories.

In my football study, defeats are exogenous so we can claim that they, not common causes, caused the market to crash. But in our setting, even showing cause and effect isn’t enough to move from data to evidence, because there are rival theories on *why* football losses cause stocks to swoon. Our hypothesis was that traders become depressed, but an alternative hypothesis is that the fall is entirely rational. Perhaps investors know that a loss will lead to employees being less productive at work and so it's sensible for them to sell stocks.

We thus ran an extra test to *support our theory*. The sentiment story requires those who trade the stock market to also care about the national team. However, many investors in the UK are international and might not care about an England defeat – some may actively celebrate it. Prior research showed that small stocks are more likely to be held domestically, since they're off the radar screen of foreign investors. Most non-Brits will have heard of AstraZeneca, a large pharmaceuticals company, but they might not know Chesnara, a small insurer. If football defeats really do affect investor mood, their impact should be even greater in small stocks, because they're particularly held by local investors – and that's what we found.

We also conducted an additional analysis to *rebut the rival theory*. If the reaction to a defeat is rational, the market should react more negatively when the defeat is unexpected. We gathered data on pre-game odds and found they had no effect on how much the stock market declines, inconsistent with the rational explanation.[##]

If you're lucky, the same test can kill both birds with one stone – support your theory while simultaneously rejecting rival theories. My employee satisfaction study controlled for common causes such as industry, size and recent performance, but I couldn't control for unobservable ones like management quality – perhaps a great manager both causes her employees to be happy and firm performance to be strong. I thus tested analyst forecasts. Investment banks like Morgan Stanley have stock analysts who predict companies' profits and give buy/sell recommendations. These analysts talk to management all the time, and so if management quality is high, they'll have optimistic earnings forecasts.

I found that the Best Companies delivered consistently higher profits than analysts predicted, suggesting that it wasn't management quality (or anything else that analysts studied) that was driving the Best Companies' success. In addition to debunking the rival theory, this test also supported my own theory: satisfied employees are more motivated, more productive and more likely to stay, ultimately boosting profits.***

In a nutshell

- *Data is evidence* if you have a randomized control trial – you arbitrarily hand out a treatment to some and a placebo to others. Since the input is *exogenous* (randomly assigned), any difference in output can be attributed to the input, and correlation is causation.
- RCTs may be expensive, and unethical if the treatment might cause harm. If so, we can't assign the input ourselves; we need to look for cases where something

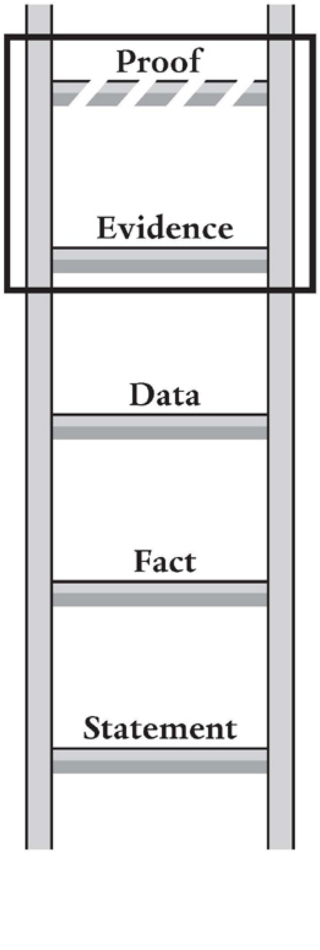
else has made the input random.

- An *instrument* shocks the input but doesn't directly affect the output. A valid instrument must be:
    - *Relevant* : it affects the input. (Rivers affect school choice.)
    - *Exogenous* : it has no effect on the output except through the input – the instrument should be somewhat crazy. (Rivers don't affect child performance.)
    - To assess whether an instrument is valid, ask: Might the instrument be correlated with the output? (Letters demanding school choice are invalid because they're written by pushy parents who also improve child performance.)
- A *natural experiment* is when an event, such as a law change, randomly divides the sample into test and control groups.
    - To assess whether a natural experiment is valid, ask: Can you affect which group you're in? Changes in investment are invalid because companies choose their level of investment.
- Valid instruments and natural experiments are like four-leaf clovers. Plan B is common sense – conduct additional tests that support your theory (for example, show that the results are larger where your hypothesis is more likely to hold) and rebut rival theories.

If you’ve succeeded in taking that tricky third step up the ladder from data to evidence, you might think you’ve reached the summit – you have *proof* of your theory. Unfortunately, you’re not yet there, and in fact you can never reach the top rung. The next chapter will explain why.

8

Evidence is Not Proof



Frederick Winslow Taylor was born in Philadelphia in 1856 to a wealthy Quaker family. Until he was twelve, he was homeschooled by his mother, a fervent abolitionist and feminist, then he spent two years studying in Europe. A promising scholar, Taylor was on track to follow in the footsteps of his father, an Ivy League-educated lawyer. He went to school at the prestigious Phillips Exeter Academy before acing the Harvard entrance exam. But due to his poor eyesight, he couldn’t take up his place at Harvard, and became a trainee machinist instead.

After finishing his apprenticeship in 1878, Taylor joined the Midvale Steel Works in Nicetown, Philadelphia. Nicetown was an ironic name, because a good chunk of Midvale’s steel was used in heavy artillery, alongside more peaceful applications such as steam turbines. Taylor started as a lathe operator but quick ly rose through the ranks, becoming machine-shop foreman and ultimately chief engineer. His rapid ascent was thanks to pioneering a radical new method to boost efficiency.

Most factory bosses saw the main productivity problem as wasted material, but Taylor viewed it as wasted effort. Some labourers didn’t know the best techniques, instead relying on tradition or rules of thumb. Others knew the tricks of the trade but had no incentive to use them as they were paid by the hour, not by output. Greater productivity would make bosses aware they could raise quotas, so they stuck to the lowest common denominator and did just enough to avoid being fired. Taylor called such behaviour ‘soldiering’, comparing it to conscripted soldiers who begrudgingly follow orders but do nothing above or beyond.

Taylor’s innovation was to transfer the tricks of the trade to management. He broke down a task into its individual parts, measured the time each required and summed everything up. Now that bosses knew how long it took to make a widget, they knew how much to pay for it. They replaced pay-by-hour with pay-by-output and introduced minimum-quantity thresholds.

This painstaking analysis achieved big productivity gains, yet Taylor wanted to go further still. He’d timed how long a task took under current methods, but what if existing techniques weren’t the most efficient? Taylor used the setting of metal cutting to take his ideas to the next level. He ran 30,000 experiments to learn the ‘one best way’ to cut metals, varying the speed, feed and shapes of cutting tools in search of the magic mixture. Then, once he’d found this one best way, he’d tell workers to follow it. Taylor wrote up his findings in a book, enticingly named *On the Art of Cutting Metals*.[1]

With such a title, it isn’t surprising the book didn’t become a bestseller. There’s nothing wrong with metal cutting, but it spoke to a limited audience. What if your factory crafted toys or weaved clothes? No matter how alluring the art of cutting metals was, you couldn’t apply the ideas to your shop floor.

So Taylor’s next book was much more general. It described how he’d successfully applied his techniques to many different fields, such as shovelling and bricklaying, as well as more mental tasks like ball-bearing inspection. He conducted experiments to find the best way to perform each task, right down to the finest detail. The optimal shovelling technique was to ‘press the forearm hard against the upper part of the right leg, just below the thigh . . . take the end of the shovel in your right hand and when you push the shovel into the pile, instead of using the muscular effort of your arms, which is tiresome, throw the weight of the body on the shovel’. This wisdom increased a shoveller’s daily output from 16 to 59 tons.

Taylor’s most vivid account described how he transformed pig-iron handling (lifting 92-pound ‘pigs’ of crude iron) at Bethlehem Steel, a company he joined in 1898. He ran tests to figure out the one best way to haul pig iron – how much to carry at a time, and how long to take breaks for – and then made his employees follow the formula. His book recounts how he explained this to a labourer called Schmidt:

> *You will do exactly as this man tells you to-morrow, from morning till night. When he tells you to pick up a pig and walk, you pick it up and you walk, and when he tells you to sit down and rest, you sit down. You do that right straight through the day. And what’s more, no back talk.*

Taylor’s programme worked, quadrupling Schmidt’s haulage from 12.5 to 47.5 tons a day. While his orders might seem like micromanagement, Taylor saw them as no different from the advice of a sports coach – as Taylor noted, ‘This is kindness; this is teaching.’ (Taylor applied his methods to his own tennis game, winning the men's doubles at the US National Championship, the precursor to the US Open.) And employees enjoyed the fruits of their productivity – Bethlehem Steel raised Schmidt's wages by 61% from $1.15 to $1.85 a day, allowing him to build a house. It was a win–win.

Taylor entitled his next book, published in 1911, *The Principles of Scientific Management*.[2] It explained how to use engineering techniques to break down a production process into individual tasks and scientifically analyse the one best way to carry out each task. Taylor claimed these principles were a universal route to riches that work in any setting – and so his book became a huge hit. *The Principles of Scientific Management* was the bestselling business book in the first half of the twentieth century, and the Academy of Management voted it the most influential management book of the entire 1900s. And its influence extended even beyond business, as we'll now see.

The boundaries of science

It wasn't just manufacturing that suffered from productivity concerns in the early-twentieth-century United States; there were similar issues in education. Seeing the success that scientific management was having in factories, some thought it could also be used to transform schools.[3] Its biggest proponent was John Franklin Bobbitt, a University of Chicago education professor who pioneered the idea of a school curriculum. The title of his article, 'The elimination of waste in education', conveys what he hoped scientific management would achieve.[4]

For Bobbitt, teachers were like factory labourers, and pupils were metals waiting to be cut into standard shapes and sizes. Teachers were too ignorant to know the best techniques, so it was up to school administrators (analogous to factory bosses) to decide on the curriculum and methods – using scientific management. They monitored classrooms to learn the range of teaching styles, trialled each approach and prescribed the one that yielded the highest scores.[5] The final step was to pay for that output. Bobbitt recommended testing kids to measure teachers' performance, which in turn determined their salaries. And the tail wagged the dog – the curriculum was designed so that it could be broken down into small parts, standardized and assessed.

Nearly a century later, the No Child Left Behind Act of 2001 took Taylorism to the next level. It linked test scores not only to teacher pay but also to federal funding and even whether a school was allowed to operate. To be financed, states had to give standardized tests to all students at certain ages. If a school failed to demonstrate 'Adequate Yearly Progress', it had to let pupils transfer to a better-performing school within the same district, and could eventually be shut down. The hope was that measurement and targets would lead to the same accountability as they did within companies. As US Secretary of Education Rod

Paige explained, ‘Good schools do operate like a business. They care about outcomes, routinely assess quality, and measure the needs of the children they serve.’[6]

With so much at stake, schools prioritized grades over education. 71% of school districts cut at least one subject to devote more time to reading and maths – with humanities, arts, social sciences, music, technology and computing on the chopping block.[7] Even among the subjects that survived the axe, teachers began teaching to the test – mechanically repeating disconnected fragments of information for kids to regurgitate in an exam, rather than showing them how to use this information and think for themselves.

Another development was the rise of scripted curriculum. Just as Taylor instructed his labourers how much to lift and when to take breaks, teachers were told what page to be on each day and which words to say. One manual prescribes how to teach children to read:

- Say *cat*. Ask: *What sound do you hear at the beginning of cat? What letter should I write in the first box?* Write *c*.
- Ask: *What sound do you hear next in cat?* Call on a child to come to the board and write *a* in the second box.
- Ask: *What sound do you hear at the end of cat? What letter should I write in the last box?* Write *t*.[8]

Teachers quit in droves, frustrated they couldn't use the style that worked best for their own personalities or their kids' learning preferences. Several went public with their resignation letters, with a forty-year veteran lamenting an over-reliance on 'data-driven education' that 'seeks only conformity, standardization, testing, and a zombie-like adherence to the shallow and generic Common Core'.[9] NCLB was criticized so much – by both Democrats and Republicans – that it was eventually replaced in 2015.

Factory reset

Why did scientific management fail in twenty-first-century education, when it had succeeded in early-twentieth-century manufacturing? Because of three key differences. A ton of iron in Philadelphia is a ton of iron in Cleveland, and shifting 20 tons is always better than 15. But in teaching, you can't compare measures of output across classrooms, no matter how much you try to standardize them. Whether a test score is good or bad depends on pupils' economic background, home life and learning challenges – teacher Yvonne could achieve an average student score of 20, yet have underperformed Zach with 15 because his students started from a lower baseline.

With Schmidt, all Taylor cared about was how much he lifts, but teaching has multifaceted output. Test scores in reading and maths are only a fraction of the value a teacher can bring – developing critical thinking, a love of learning and a respect for different viewpoints. Finally, while there might be one best method to cut metals, the most effective way to run a class depends on the teacher and her pupils, and it's her that's best placed to figure out what works. The biggest waste in education isn't teachers slacking off but their being prevented from using their initiative.

These differences are so stark that they should have been obvious. But the supporters of scientific management in education failed to realize that *evidence is not proof*, because it may not be *universal*. A proof is absolute. When Archimedes showed that the area of a circle is pi times the square of its radius, he proved this not just for circles in ancient Greece in the third century BCE but also for circles across the world today. Evidence, in contrast, might only apply to the context in which it was gathered. Finding that scientific management works in pig-iron hauling, shovelling and metal cutting doesn't mean it succeeds in schools.[*]

The usefulness of evidence depends on its *validity*. Chapters 5–7 focused on *internal validity* – whether it actually finds what it claims in the context studied. When people are able to activate their System 2 and scrutinize evidence, they typically assess this aspect of validity. Often overlooked is *external validity* – even if a study has perfect internal validity in one context, it may not apply to different settings.

We often mistake evidence for proof due to the twin biases. Confirmation bias kicks in if we're wired to think a result holds and then naïvely accept studies that demonstrate this result even in a different context. Secretary of Education Rod Paige initially served in the Navy and then became an American football coach. These are fields in which there are many ‘one best ways’ – how to throw a football spiral, for example – so he was prone to think they’re similarly plentiful in education. Black-and-white thinking means we believe a practice always works, or always backfires, even though the world is granular. If scientific management flourishes in factories, people think it will also succeed in schools.

These biases mean there’s a simple blueprint to write a viral article or bestselling book. You claim universality – a ‘theory of everything’ that applies in all settings. Then you sit back and wait for everyone to buy it, regardless of the company they run, the vocation they pursue or the dreams they chase.

To buttress this ‘theory of everything’, you have two options. One is to find success stories across multiple fields and highlight how your theory is the one common factor behind them, so it must explain them all. Simon Sinek goes further than claiming that *why* explains Apple’s success – he pronounces it a universal route to stardom, explaining how the Wright Brothers beat the wealthy Samuel Pierpont Langley to launch the first flight, and how volunteer-run Wikipedia surpassed Microsoftbacked Encarta to become the world’s main resource for knowledge. Apple, the Wright Brothers and Wikipedia are all in very different areas, so ‘starting with *why* ’ must be their magic formula, as they all share it.

By now, we can probably spot all the problems. The sample is selected (it doesn’t contain those with a *why* who failed), there’s no control group (those without a *why* who succeeded), and there’s no consideration of alternative explanations (the trio have many other things in common). Yet you can cheerfully ignore internal validity if confirmation bias means that people like your theory, and black-and-white thinking means they believe there’s a single unifying explanation for success in every context. *The Spirit Level* claimed that inequality is the cause of all the ills of the world, ignoring rival theories such as poverty. The *Guardian* was reeled in, endorsing the book’s front cover with ‘A sweeping theory of everything’.

The second path to stardom is the opposite. You find the one setting where your results are the strongest: in pig-iron handling, scientific management didn’t just boost output, it quadrupled it. You might even ensure internal validity through a correct control group. Then you claim these results apply everywhere, this time cheerfully ignoring external validity. A book might promise a magic formula to lose weight with glowing testimonials from people for whom it genuinely worked. Yet it might not succeed for others. There may be no one best way to shed pounds, because it depends on your fitness, age and time available for exercise. But a book focused on working mothers in their forties will have a far smaller audience than one that promises a ‘four-hour body’ to all – particularly if the latter boasts the subtitle ‘An Uncommon Guide to Rapid Fat-loss, Incredible Sex, and Becoming Superhuman’ so that readers fall for it hook, line and sinker.[10]

The opening sentences of *Built to Last* are ‘We believe every CEO, manager, and entrepreneur in the world should read this book. So should every board member, consultant, investor, journalist, business student.’ It goes beyond companies; they claim that the book applies ‘to school districts, colleges, universities, churches, teams, governments, and even families and individuals’. And it’s wider than the US – ‘the central concepts in *Built to Last* apply worldwide, across cultures and in multicultural environments’. Claiming universality is a smart way to rack up sales, but as we saw in Chapter 4 this book had no internal validity – and its external validity is no better. Even if the authors had fully nailed what drives the success of eighteen large publicly traded companies in the US, this might not be applicable to start-ups, nor to non-profits, nor to companies in other countries.

How do we deal with granularity? The ideal approach is to find a study on the precise context you’re interested in. Emily Oster didn’t just assume that the study on the link between breastfeeding and IQ applied also to infections, diabetes and obesity; she meticulously searched for papers on these other outcomes. But that’s often impossible. No study will be replicated everywhere, as it’s costly to test every sector in every country – there’s no analysis of how the *Built to Last* principles affect Brazilian churches.

There’s also a trade-off between internal and external validity – the research that’s most reliable might not be in the closest possible context. Instruments and natural experiments allow us to demonstrate causation, but they’re difficult to come by and are only available in narrow settings. Caroline Hoxby used rivers to show how child performance is affected by the number of school districts, but since rivers don’t affect charter schools or academies, they can’t shed light on these other ways to increase school choice.

All this means that interpreting data comes down to common sense, not just statistical pyrotechnics. A simple question to ask is: *Are there any reasons why the finding might not apply to our context?* The first study showing that smoking causes cancer was conducted in the US in 1950. These findings likely apply across the globe and to the present day. The effect of smoking on cancer is related to human physiology, which is broadly similar throughout the world and has changed little between 1950 and today.

But whether scientific management is the right tool hinges on the context of the job. Some tasks have a clear best method and a single measurable output, others accommodate different approaches and have multiple goals; some employees appreciate detailed instructions, others feel micromanaged.

Yet what if the context you're interested in is the same as the context of the study – surely then the study's insights can be applied? Unfortunately not, as we'll now see.

Grit and greatness

Each year, around 1,200 new cadets are accepted into the US Military Academy at West Point, New York. As they report for duty on their first day, most feel mixed emotions – pride about serving their country, but anxiety about whether they'll survive 'Beast Barracks'. This gruelling orientation and training course involves seventeen-hour days and lasts six weeks over the summer. The objective: to transition men and women from civilian to military life.

Hearts pound, muscles ache and lungs burn. A common phrase at West Point is that 'Every cadet is an athlete' – a sentiment reflected in the intense physical regimen that new recruits face. Starting before sunrise and ending after dusk, cadets are put through a relentless barrage of endurance and strength challenges. In between the physical exertions, they're kept on their toes with mental exercises and classroom lectures. The 'celebration' to commemorate the end of the six weeks is a ruck march. Starting at 2.30 a.m., the recruits trek twelve miles in full gear, carrying a 40-pound load of equipment on their backs over steep and rocky hills back to the West Point main campus. It's a final test of body and mind, representing all they've faced and conquered during the programme.

Every single day, cadets feel they've reached breaking point and are tempted to quit. They know their dreams of a military career will be dashed, but the mental exhaustion and physical pain are just too great for some, and up to one in five will throw in the towel.[11] Beast Barracks certainly isn't for the fainthearted. But do certain personal attributes make success more likely? Would these same characteristics give people an edge in other fields of work, education and life?

Twenty-seven-year-old management consultant Angela Duckworth left her high-pressure job for one that was even more demanding – teaching maths to seventh-grade students in a New York City public school. As she graded papers and marked quizzes, Duckworth noticed an intriguing pattern. Her top-performing students were often not the smartest kids in the room. She concluded that IQ, which many parents obsess over and measure relentlessly, may not be the true key to high achievement.

Duckworth wanted to explore what really drives success, and so she embarked on a Ph.D. in Psychology at the University of Pennsylvania. Her work took her to West Point, where she and fellow researchers were invited to conduct a study.

West Point's leaders were interested in predicting which new recruits would complete Beast Barracks. They thought the biggest factor would be the Whole Candidate Score (WCS), a measure West Point tailor-made to capture the full range of qualities they believed were essential for a recruit

– academic ability, leadership and physical fitness – and used as the number-one criterion for admission.

But Duckworth and her team found that the WCS wasn't a good predictor for whether a candidate survived Beast Barracks. They had an alternative hypothesis: what matters isn't raw mental and physical ability, but grit, which they defined as a combination of passion and perseverance.[†] In 2004, they designed a survey to assess grit and gave it to the new cadets at the start of that year's Beast Barracks.[12] Recruits were asked to rate how closely twelve different statements applied to them. Some were about work ethic ('I am a hard worker'; 'I am diligent'); others explored persistence or singular focus ('I often set a goal but later choose to pursue a different one'; 'My interests change from year to year').

Duckworth and her colleagues found that grit scores significantly predicted Beast graduation. They took many of the steps we've highlighted in this book. They studied a representative sample of all the cadets who started Beast, rather than a selected sample of only those who crossed the finishing line. They had a clear control group, comparing gritty to nongritty recruits. They also controlled for common causes, showing that grit mattered, even holding the WCS constant. And, by conducting the survey at the start, they ruled out reverse causation (cadets who survived Beast then reported themselves as being gritty).

Duckworth was also aware of external validity. Grit may matter for a physical challenge like Beast Barracks, but in more intellectual contexts, surely it would play second fiddle to IQ? Her team thus studied different settings. They found that grit predicted how far kids progressed in the National Spelling Bee finals, even when controlling for IQ.[13] Grit also forecast the grades of 139 University of Pennsylvania undergrads, over and above SAT scores.[‡][14] It seemed they'd found a 'theory of everything' – grit matters in all settings. Duckworth's TED talk, 'Grit: the power of passion and perseverance', claims that 'in all those very different contexts, one characteristic emerged as a significant predictor of success . . . grit'. Her book, *Grit*, makes similar assertions, and in a *New York Times* interview to promote it, she declared that grit 'beats the pants off IQ, SAT scores, physical fitness and a bazillion other measures'.[15]

Such statements feed on our biases. Anyone can develop grit, whereas physical prowess is partly genetic, so Duckworth's message is empowering. It's no surprise that Duckworth's TED talk has been viewed 30 million times and her book was a *New York Times* bestseller. But isn't this interest fully deserved? Given that she's claimed to have addressed all the concerns with statements, facts, data and evidence, shouldn't I stop being a killjoy and acknowledge that biases might lead us to accept something that's correct?[16]

Not necessarily. The limitation of all the above studies is *restriction of range*. Due to moderation, something may be good or bad only up to a point, so what you see in a narrow range of values isn’t true in general. Duckworth studied cadets who’d already got into West Point. They were extremely fit to begin with, and there are diminishing returns to fitness after a certain level – if you can already sprint 100 metres in twelve seconds, shaving off an extra second is unlikely to help you complete Beast. Since fitness wasn’t that relevant, other factors like grit mattered instead. However, for the average teenager aspiring to join the military, she might be better off improving her fitness rather than bolstering her grit.

Similarly, the students in the University of Pennsylvania study already had SAT scores in the 96th percentile, but for the average undergraduate, SAT scores might be a strong predictor of grades. Kids in the Spelling Bee finals were already in the highest echelon of IQ. Beyond those settings, for mere mortals wanting to become a successful writer, singer or doctor, it may be even more important to work on honing your craft than your grit.

The problem here isn’t over-extrapolation to a different context (ignoring granularity), but over-extrapolation to a different range (ignoring moderation). Even if you’re interested in the US military rather than another country or profession, what you care about is how the average person succeeds – not someone who’s already good enough to get into West Point.

Note that the way in which moderation rears its ugly head is different from the examples in Chapter 1. There, we discussed over-extrapolation to different levels of the *input*.

Studies show that increasing water intake improves athletic performance, but only up to the level needed to prevent dehydration.[17] The articles that recommended drinking as much as possible overextrapolated outside that range, leading David Rogers to consume far more water than was tested in those studies. Here, the issue is over-extrapolation to different levels of the *control*. We're interested in the grit, not the controls (IQ, SAT scores or WCS numbers). However, if the controls are already so high that they're now irrelevant, then we overestimate the importance of the input of interest: grit.

What if we're only interested in showing that grit is important, rather than claiming that it's *more* important than IQ? Unfortunately, restriction of range remains a problem. It may be that grit only matters when your ability is also high. This is known as an *interaction effect* – grit doesn't help as a standalone, but only when combined with, or interacted with, ability. It didn't matter how much passion or perseverance William Hung had – with limited musical talent, he'd never win *American Idol*. By focusing on successful athletes, students and spellers, Duckworth only demonstrated that grit mattered when accompanied by very high ability; it may be irrelevant for the average Joe.

The best response to restriction of range is – you guessed it – common sense. The first step is to be crystal clear about what was actually studied. Headlines stress how 'a study finds that X improves Y'. But what were the ranges of

X investigated? Showing that increasing water intake from 1 to 2 litres improves marathon performance may not mean that raising it from 2 to 3 will have similar effects. Besides the input you're interested in, were the subjects special in other dimensions, such as being intimidatingly fit or having scarily high IQs, that might cause the input to have a different effect to how it affects the general population?

Then, if the research doesn't cover the range we're interested in, we can use common sense to consider whether the results might still apply. If studies on smoking show that 1 cigarette a day harms your health more than 0, and 2 more than 1, all the way up to 50, there's no obvious rationale why puffing more than 50 won't continue to cause damage. For water intake, there's a logical reason why you can have too much – it dilutes your essential minerals. Even for something that seems good in itself, such as sleep, splurging on it reduces the time you have for exercise, socializing and family.

The parachute that didn't work

In December 2018, a study on the ineffectiveness of parachutes hit national headlines.[18] Harvard Medical School cardiologist Robert Yeh led a research team that found twenty-three volunteers willing to jump out of either a biplane in Massachusetts or a helicopter in Michigan. It was the perfect example of a randomized control trial. Half were randomly assigned a fully working parachute, the other half an empty backpack. Amazingly, the researchers found *no difference* in the injury rate of the two groups.

The study was published in the prestigious *British Medical Journal* and obtained as close to causation as you can get. So why haven't its results caught on – why are skydivers still using parachutes? It's not immediately clear why external validity might be limited. Sure, the tests were only carried out in Massachusetts and Michigan, but there's no obvious reason why the results will be different elsewhere. And studies using biplanes and helicopters likely also apply to other aircraft.

But one detail was crucial. Both the biplane and the helicopter were parked on the ground, so the volunteers only jumped two feet. The range of the fall was restricted, and so the paper had no implications for larger jumps. It was a satirical study, warning about the danger of researchers cherry-picking settings where they get the result they want, and then over-extrapolating it to other contexts. As the authors wryly note in their final sentence, 'although we can confidently recommend that individuals jumping from small stationary aircraft on the ground do not require parachutes, individual judgement should be exercised when applying these findings at higher altitudes'. Whether you're skydiving, dreaming of the military or crafting a nation's education policy, remember that context and range are key.

In a nutshell

- *Evidence is not proof* because it may not be *universal*. Even if evidence has *internal validity* (uncovers causation), it may not have *external validity* (apply in different settings).
- External validity may be absent due to *granularity*. Evidence that a practice works in one profession, industry or country may not be generalizable to others.
    - If you can’t find a study on the precise setting that you’re interested in, ask: Are there any reasons why the finding might not apply to your context?
- External validity may also be absent due to *moderation*.
    - There may be moderation in the input. Showing that 2 litres of water per hour are better than 1 doesn’t mean that 3 are better than 2.
    - There may be moderation in the controls. Showing that grit matters more than fitness for West Point recruits, who are already very fit, needn’t imply that it matters more for the general population.
    - The inputs may interact with the controls. Showing that grit matters (at all, not just more than fitness) for West Point recruits needn’t imply that it matters for the general population. Grit might only matter if you’re very fit.
    - Ask: What were the ranges of the input and controls

studied? Are there reasons why the results might not hold for the ranges we’re interested in?

We’ve now reached the end of Part II, which has explained how to avoid missteps up the Ladder of Misinference. But to understand the world better and make shrewder decisions, we need to do more than just interpret statements, facts, data and evidence correctly. In our everyday lives, we’ll encounter information and form our opinions from less formal sources. As individuals, we gather intelligence from books, newspapers and our friends and colleagues. In organizations, we aim to pool our knowledge, hoping that a patchwork of perspectives will form a coherent whole. As a society, our collective understanding is shaped by public messaging, social media and our school curriculums. Part III will explore how we can think smarter more broadly – as individuals, organizations and societies.

PART III

The Solutions

9

Thinking Smarter as Individuals

My time at Merton College, Oxford University, had been a happy one. I emigrated to the US two years after graduating but retained strong ties and served as my year's class secretary, gathering alumni news and geeing up attendance at events. After returning to London a decade later, I was a regular at our annual Merton in the City reunion. So when an email headed 'Merton in the City, 2016, Tues 2 Feb' landed in my inbox, I instinctively moved my cursor to the 'Book now' link.

Just as I was about to click, I saw the blurb: 'This year's speaker will be Roger Bootle (1970) giving a talk entitled "To stay or go? – the EU and the UK" . . . Roger is the founder and Managing Director of Capital Economics . . . He has also been Group Chief Economist of the HSBC Group.'

This made it even more of a no-brainer. With the Brexit vote just four months away, they couldn't have chosen a more topical issue. The bubble I'd grown up in and now worked in meant that virtually all my LinkedIn, X and Facebook contacts were Remainers. They often tagged me in posts asking how to rebut a Brexit argument or explain a justification for Remain in more detail. This event would give me even more ammunition to fight the good fight.

But the next line of the email caught my eye: 'He is the author of six books, including *The Trouble with Europe*.' The trouble with Europe? Surely not! Yet a quick web search for Roger Bootle confirmed that he's a Eurosceptic. When I first saw the event was on the EU, it never occurred to me that the speaker might support Leave. He went to Merton like me, was an economist like me, and worked in cosmopolitan London like me ('Capital' Economics), so he must be a Remainer like me. It crossed my mind to skip it – telling myself this wasn't because of biased search, but because Brexiters' views came from the side of a bus and I shouldn't support an event that spread misinformation. Yet with free bubbly on the line, my lofty principles went to the wall. I hit 'Book now'.

When the evening rolled around, I was surprised to find that all of Roger's points were well grounded and logically argued. Even though I didn't agree with some of them, I could at least see where he was coming from. To this day it remains the most eye-opening talk I've ever attended. When I got home, I was eager to learn more – I combed through Capital Economics' report on Brexit, cross-checked a few points and delved deeper into others. I then wrote up the main arguments in a post on my blog, titled ‘The case for Brexit’.

This book has so far focused on how we can avoid being misled. But becoming more knowledgeable and making better decisions isn’t just about defending against misinformation – it’s also about positively gathering information. The above episode highlights the power of *actively seeking dissenting viewpoints*. In a trial, the judge ensures the jury hears evidence on both sides. Similarly, we’ll have the best chance of getting highstakes decisions right by seeing the case for the defence as well as that for the prosecution.

For a well-defined issue like Brexit, it’s as simple as reading articles that take the opposite position. Doing so isn’t betraying our ideals; as is commonly attributed to Aristotle, ‘It is the mark of an educated mind to be able to entertain a thought without accepting it.’ Even if we think that 90% of what they say is wrong, 10% might be right, and this 10% means we come away smarter than we were beforehand.

It’s crazy how one of the biggest compliments you can give an article is ‘That’s exactly what I wanted to say but you said it better.’ If that’s the case, you learned nothing from it – other than the art of rhetoric. In contrast, the greatest snub is to unfollow someone because we don’t like what they’ve posted; we think we’re punishing them, but we’re hurting ourselves by removing opportunities to learn.[*] As a sustainability advocate, I want to read every well-informed critique that's out there. It's not so much about being open-minded as self-interested. If I'm aware of the main counterarguments, I can include them in any talk I give – either to acknowledge them, making for a more balanced and thus persuasive speech, or to rebut them pre-emptively.

For sustainability, I know who the biggest sceptics are, so I can be on the lookout for their latest thoughts. I also subscribe to a daily newsletter that compiles the most recent articles on the subject; I don't have the capacity to read even half of them, but I'll prioritize any anti-sustainability content. For other topics, we might have no idea who the dissenters are, but we can simply google the opposite of what we'd like to be true. A coffee addict could enter 'why caffeine is bad for you' and see whether it throws out any high-quality evidence.

Beyond specific subjects, we can use this approach to broaden our world view more generally, for instance by signing up to newspapers or writers whose stances disagree with ours. I'm subscribed to both the right-wing *Telegraph* and the left-wing *Guardian*. This means that, on any topical issue – abortion, immigration or the legalization of drugs – I'll hopefully have seen both sides.

Reading dissenting articles is only the starting point. We might do so to tick a box and delude ourselves we're openminded, but we go through them with the mindset of trying to tear them apart rather than learn. As business professor Stephen Covey commented, 'Most people do not listen with the intent to understand; they listen with the intent to reply.'[1]

Back in Chapter 1, we saw the Bayesian inference diagram below. The bottom mentions another term, but we never discussed it back then.

Does *Information* support *Hypothesis?*

Depends on

Is *Information* consistent with *Hypothesis?*

vs

Is *Information* consistent with *Alternative Hypotheses?*

(plus another term)

That other term is the strength of your initial belief. Evidence allows you to learn whether your hypothesis is true, but learning is always relative to a starting point.[†] If you go in with absolute certainty, the mathematics behind Bayesian inference shows – indeed, proves – that your beliefs can never change, regardless of any new information that comes to light. Only if you're truly open to the possibility of being wrong can you learn. As Leo Tolstoy wrote, 'The most difficult subjects can be explained to the most slow-witted man if he has not formed any idea of them already; but the simplest thing cannot be made clear to the most intelligent man if he is firmly persuaded that he knows already.'

Standing on the shoulders of giants

We've highlighted the value of reading articles on the other side, but how do we know which ones to trust? Part II provided some simple questions to check if a conclusion is valid, but this still takes effort. Is there an easy way of knowing whether a study is worth our time?

There is, thanks to the peer-review process. Peer review might sound like an arcane scholarly ritual, irrelevant to anyone outside the ivory tower, but it's as valuable as any other certification mechanism. We trust that a medicine is safe to swallow if the US Food and Drug Administration has approved it; we have faith in a company's accounts if a reputable auditor has signed off on them; we sleep soundly at night if our locks bear the British Standards Institute Kitemark. Peer review does the same for academic studies. (We'll later tackle other sources of information such as books and newspaper articles).

When a paper is submitted to a scientific journal, the editor asks other leading scholars for their confidential views on its quality. The standards can be dauntingly high, with the most elite journals rejecting up to 95% of submissions. The remaining 5% aren't immediately accepted either; instead, their status is 'revise and resubmit'. The editor and reviewers highlight concerns for the authors to address, and the paper can still be rejected at the next round. It's not unusual for a manuscript to take five years to be published after its first draft – a hard slog for the authors, but essential to give readers confidence in the results. As discussed in the Introduction, a study on pay gaps did a 180 on its initial conclusion after the review process forced it to correct its mistakes.

Peer review allows us to stand on the shoulders of giants – leading minds who've already probed a paper for us. In contrast, we should be sceptical of unvetted research, because its claims might not be supported by the evidence. Sometimes it's not even close, like the diversity article whose punchline was contradicted by all ninety of its tests. If a study hasn't been scrutinized, it can claim anything.

We can quickly check whether an article has been peer-reviewed by seeing if it's published in a scientific journal, like *Nature*, the *British Medical Journal* or the *Journal of Finance*. If it's simply posted on an author's webpage or a company's website, then it hasn't been certified and we should treat it with caution. But just appearing in *a* journal doesn't mean much – we need to ask *which* journal, because there's a vast range in reviewing standards. The analytics company Cabell's classifies over 15,000 journals as 'predatory', for transgressions such as claiming to peer-review their papers when they actually don't. Journal quality can easily be verified: for business, the *Financial*

Times compiles a Top 50 list; Scimago has rankings across all fields.

But does peer review matter for the real world? Isn't it just academics squabbling among themselves, boasting that their paper was peer-reviewed and their nemesis's wasn't? Sure, research results may get overturned, but who cares if the government forces companies to disclose their pay gap based on shoddy data?

It does matter. Disclosure costs US companies an estimated $1.3 billion in the first year and $530 million each subsequent year.[2] That's only the cost of gathering the information, not how it might distort decisions – for example, companies replacing low-paid workers with machines to reduce their pay gap. And it matters in far more important settings than pay gaps, as we'll now see.

'Information about Theranos, a privately held biotechnology company that has developed novel approaches for laboratory diagnostic testing, has appeared in the *Wall Street Journal*, *Business Insider*, *San Francisco Business Times*, *Fortune*, *Forbes*, *Medscape* and *Silicon Valley Business Journal* – but not in the peer-reviewed biomedical literature.'

That's the opening sentence of an article by Stanford medicine professor John Ioannidis, published in the *Journal of the American Medical Association* in February 2015.[3] It highlighted that none of Theranos's claims had ever been vetted by other scientists.

But no one cared. Elizabeth Holmes had puffed her promises so powerfully, and the media had hyped her so heatedly, that people lapped up whatever she said – evidence or no evidence. At the time of Ioannidis's article, Theranos was worth a staggering $9 billion. Eight months later, *Wall Street Journal* writer John Carreyrou exposed Theranos's fraud in a series of columns that formed the basis for his 2018 bestseller *Bad Blood*.[4] Carreyrou's work saved thousands, perhaps millions, of patients, employees and investors from being fooled by Theranos. Yet it wouldn't have been necessary had a few key people recognized the value of peer review and not taken Theranos's claims at face value.

Finding common ground

Certification can never be perfect. The accounts of energy company Enron had been signed off by Arthur Andersen, then one of the world's leading auditors, but were later exposed as fraudulent. The arthritis medicine Vioxx was approved in 1999 but withdrawn five years later due to links with heart attacks and strokes. The same is true for academic journals – no matter how diligent they are, reviewers and editors can't spot every flaw. For example, it's difficult to detect data mining because they never see the tests the authors tried and buried because they didn't work out.

Journals can also fall victim to *publication bias*, where they accept a paper because they like its findings. And what findings do editors like? Statistically significant ones, because they're more likely to make a splash. The main measure of a journal's reputation is its 'impact factor', the number of times its papers are cited by other journals, and people are more likely to quote a study that finds something than one that unearths nothing.

An ingenious randomized control trial put this to the test. It took two versions of the same manuscript, with everything identical except that one had significant results and the other didn't, and submitted them to scientific journals in orthopaedics.[5] Reviewers were more likely to recommend the former for publication, as well as to detect errors in the latter's methodology – even though it was exactly the same.

What does this mean for the reliability of peer review? For unpublished papers, it doesn't change anything – we should still be sceptical. But it does mean that published papers can't be viewed as gospel; publication increases our confidence that it's accurate, but not all the way to 100%. Yet perfect shouldn't be the enemy of good. Even though verification isn't absolute, it's better to go with something checked than unchecked. In addition, the process is usually self-correcting for the worst mistakes. Scientists have strong incentives to disprove influential papers because it can bring them fame, and journals will publish debunking studies – even studies that overturn their own publications – to maintain their reputation for accuracy.

A famous paper co-authored by Amy Cuddy suggested that you can ace job interviews and public speeches by ‘power posing’ beforehand, like standing with your arms outstretched in a victory celebration. It became the basis for the second-most-viewed TED talk in history, ‘Your body language shapes who you are’.[6] But *Psychological Science*, the same journal that published Cuddy’s research, later released a paper by other authors who conducted a similar experiment[‡] but found no effects.[7] The *Lancet*, which featured Andrew Wakefield’s study linking vaccination to autism,[8] went a step further and subsequently retracted it.[§] In these cases, a quick internet search will show us if a paper has been retracted or overturned – the Wikipedia entries for ‘power posing’, ‘Amy Cuddy’, ‘*Lancet* MMR autism fraud’ and ‘Andrew Wakefield’ contain all the gory details.

While retractions and debunkings attract the greatest attention, they’re not the main way science evolves. Instead, a paper’s results may get overturned not due to carelessness, but because new methodologies develop or better datasets become available. Sometimes, no method is clearly best and different research teams legitimately reach contrasting conclusions.[||] Combined with peer review being imperfect, this means that we should avoid putting too much weight on a single study. Instead, we should seek out the *scientific consensus* on a topic.

The best way to do so is to read a *systematic review*, sometimes known as a *survey paper* or *review paper*, such as the Cochrane reviews for medicine. The review editors commission the leading authorities on a topic to write an overview that captures the consensus of the scientific community, draws out the points of agreement and highlights the areas that are in dispute or not yet explored. We can think of them like a Wikipedia entry, but more in depth and on more specialist subjects – and, most importantly, only written by experts. Public bodies often conduct systematic reviews: for example, Australia's National Health and Medical Research Council published one on the ineffectiveness of homeopathy.[9] Even more accessible are websites such as that of the UK's National Health Service, which summarize the scientific consensus on how to prevent and treat illnesses and whether a medicine is safe and effective.

A systematic review will also convey the volume of evidence on each side. If the overwhelming majority of papers find that climate change is predominantly man-made, with very few documenting the opposite, the review will communicate this. In contrast, if every mention of the 10,000-hours rule refers to Malcolm Gladwell, then you only have a single source. Just as importantly, a review puts greater weight on more rigorous papers rather than simply counting the number of Ayes and Nays. Cochrane prioritizes randomized control trials; Emily Oster focused on breastfeeding studies that controlled for common causes – those that contained evidence, not just data.

From Harvard Yard to Fleet Street

Most of us won't spend our Sunday afternoons trawling through academic journals. Fortunately, there's a wealth of information beyond scientific papers. Books, newspapers, and magazines like *New Scientist*, *National Geographic* and *Fortune* make complex studies accessible to a general audience. Companies, regulators and NGOs release reports but never try to publish them in journals because that's not their goal. While some of these articles are unreliable, as we've seen, others may have significant merit.

How do we evaluate these sources? The first step is to *treat them with caution* – like anything else that hasn't been certified. They still may be valid, but they don't have the seal of external approval. Often a book implies credibility – doing something 'by the book' suggests you're following best practice; if you 'wrote the book' you're seen as an authority – but it's subject to none of the review process of an academic study.[#] 'I wrote the book on X' is no more authoritative than 'I wrote a series of blogs on X' or 'I've given YouTube tutorials on X.'

The second step is to *ask an expert*. We might personally know one; if we don't, we saw in Chapter 3 how we can google phrases such as 'Why we sleep criticism' to find an informed opinion. If we learn about research through a newspaper column, it's more credible if the journalist has included views on the study from experts who've published their own papers on the topic and don't have a vested interest in the study being true.

We can also *assess whether an article is balanced*. Do the authors acknowledge alternative explanations, or caution that their results may not apply in other settings? In social sciences, evidence is almost never proof, so claims to have found 'clear evidence' or proven something 'beyond doubt' are a giveaway that the researchers didn't seriously consider rival theories.[**] The accountancy firm Grant Thornton released a study with the subtitle 'A proven link between effective corporate governance and value creation'.[10] The foreword declared 'there has never been conclusive proof . . . until now' and suggested they'd found 'the holy grail'. Yet even a cursory glance uncovers reverse causation and common causes.

Similarly, some authors exaggerate how radical their results are. The authors of *Built to Last* stress how their findings were so revolutionary that they nearly fainted: 'much of what we found surprised us – even shocked us at times. Widely held myths fell by the dozen. Traditional frameworks buckled and cracked. Midway through the project, we found ourselves disoriented, as evidence flew in the face of many of our own preconceptions.' But it's not for authors to call their findings ground-breaking – that's for the reader to judge.

In *Hamlet*, a tell-tale sign of deception was when a character 'doth protest too much'.[††] If you need to shout about the conclusiveness of your proof or the novelty of your results, maybe they're not strong enough to speak for themselves.

Play the (wo)man, not just the ball

Another guardrail is to consider who the authors are. On the face of it, this seems ad hominem – playing the man, not the ball. But just as a trial considers the credibility of an expert witness, ad hominem assessments *are* valid if they're focused on details relevant for the person's reliability.

Two factors are relevant: bias and credentials. Starting with bias, organizations may have existing positions that they want to defend. Any report on CEO pay by the High Pay Centre will conclude that CEOs are highly paid. Others may have products that the research helps push. Grant Thornton's measure of corporate governance was a company's score on the Grant Thornton Corporate Governance Index, so it's not surprising they found that this score matters. Others still may enjoy a PR boost from the claims. McKinsey became known as a beacon for long-term thinking after their study touting its benefits. Academics are also prone to bias, particularly those famous for a position – if they're known for wielding a hammer, they'll see everything as a nail. For example, *The Spirit Level* authors seem eager to believe that inequality is the source of every single problem in the world.

A clever paper by Brian Fabo and co-authors inspected these biases systematically. Called 'Fifty shades of QE', it investigated fifty-four studies on the effect of quantitative easing – central banks buying government bonds – which became popular following the 2007–8 financial crisis.[11] Some of these studies were written by economists working for central banks, others by academics at universities. The authors found that papers by central bank economists claimed much more positive effects of QE than those by academics. To paraphrase Mandy Rice-Davies again, 'They would say that, wouldn't they?' The researchers also discovered that these economists were rewarded with promotions, perhaps as a thank-you for justifying their employers' policies.

For all these sources of bias, we can ask ourselves: *What are the authors' incentives to claim their result?* In all the above cases, they benefited from doing so. This doesn't mean the studies are incorrect, but it does mean we should view them with healthy scepticism. This question applies beyond studies to stories and statements – Theranos received a $9 billion valuation based on its promises, and Tariq Fancy became a high-priced keynote speaker off the back of his allegations about sustainable investing. In contrast, Galileo Galilei was imprisoned for claiming that the sun, not the Earth, is the centre of the universe; he had no incentive except for the pursuit of the truth.

A starker question is: *Would the authors have published the paper if it had found the opposite result?* The High Pay Centre won't acknowledge that CEOs are underpaid, nor would Grant Thornton admit that its index was irrelevant for firm performance. Yet some academics have released studies with unpopular findings, such as higher pay gaps being associated with greater performance.

Turning to credentials, they stem from three sources for a scientific study. The first is an author's *research qualifications*. A Ph.D. is a basic certification to conduct academic research, just like you'd hope for a dentistry degree before letting someone drill your teeth. Yet an influential author who self-styles as 'one of Britain's leading economists' – beware the superlative – and gives himself the title 'Professor' isn't actually a professor anywhere, nor even a doctor, as he has no Ph.D.;[‡‡] he hasn't published any articles in even third-tier journals. In other cases, people call themselves Professor or Doctor – sometimes highlighting it in their LinkedIn profile name, X handle or book cover – when they're an adjunct professor or have an honorary doctorate.[§§] Such people often have significant practitioner expertise, but that's different from the ability to conduct scientific research.

A Ph.D. isn't everything; some go on to win Nobel Prizes while others tread water, so we need to look further. A second source of credentials is the author's *track record* of top-tier publications in the relevant field. Nearly all academics make their publications available on their website; if they don’t, that’s a negative signal. A third is the quality of their *institution*. This isn’t elitism but simply a desire to use the best evidence. We’d listen more closely to a medical opinion from the Royal Marsden Hospital than one we’ve never heard of. In contrast, I’ve seen many organizations mention ‘research by the University of Sunnybeach’ because it supports their position, when they’d never hire anyone from the University of Sunnybeach.

It’s important to take into account *both* the authors and their institution. Often people refer to ‘a study by Harvard University’. This makes no sense, because Harvard University doesn’t release studies. Anyone with the loosest affiliation to Harvard can post a paper without Harvard’s approval – a tenured professor, a part-time lecturer with no research duties or a master’s student writing a thesis. (This contrasts with organizations which do require institutional sign-off, so it makes sense to refer to a McKinsey study.) The author should be mentioned in addition to the institution, such as ‘a study by Professor Terry Odean of Berkeley’.

None of this is to say that well-published authors at top institutions are always right and others are always wrong. Credentials are simply *one* factor to assess when evaluating evidence, just as a company’s brand name is one input into a purchasing decision. We should ask a similar question to the one used to assess bias: *If the same study was written by the same authors, with the same credentials, but found the opposite results, would you still believe it?*

Author credentials are equally relevant for books. For those written by academics, we can ask the same questions as above. Journalists are also frequent book authors, given their skill in telling engaging stories and synthesizing a vast array of findings into memorable themes. Their relevant credentials are the newspaper they write for and their past articles on the topic, to ensure they're not jumping on something hot without expertise.

Practitioners, such as CEOs and investors, have substantial experience that they can draw from. There's nobody more qualified to write an account of the companies they've run or the investments they've made. However, some move beyond telling war stories to proclaiming a universal set of rules for success. Even if they could overcome the narrative fallacy and perfectly identify what led to their accomplishments, without scientific research we don't know whether these principles work in general.

But many other books, particularly in the self-help genre, are written by authors with neither proficiency in producing or synthesizing evidence, nor expertise from leading companies. A popular book on time management, *Getting Things Done*,[12] is penned by David Allen, a former landscaper, vitamin distributor, glass blower, travel agent, petrol station manager, U-Haul dealer, moped salesman and chef.[13] It's sold over 1.5 million copies, but it's not clear what Allen’s expertise is based on. The book contains almost no references; instead his assertions often start with ‘in my experience’, as if that were proof.

Especially in self-help, it’s easy to become influential if your advice plays into the twin biases. A business magazine noted that ‘Fortunately for Allen, he didn’t need empirical evidence: People felt better after taking his seminars,’ and continued: ‘No studies exist proving that [Allen’s methodology] increases productivity, decreases stress, or boosts the bottom line, Allen admits.’ Similarly, Simon Sinek is an ex-advertising salesman, not a neuroscientist with knowledge of the limbic brain, nor a business professor who’s tested the hypothesis that having a *why* leads to success, nor a CEO who can draw from decades of experience, albeit at specific companies.

Just as weak papers often exaggerate their conclusions, writers do so with their credentials. Claims to be a ‘bestselling author’ are often meaningless, as there’s no clear definition of this status – whether you need to be in the top ten or the top thousand, whether for all books or just a small subfield, or for how long (Amazon’s list is updated every hour).[‖‖] ‘One of Britain’s leading economists’ is equally unverifiable, as there’s no clear ranking system,[##] and ‘global influencer’ or ‘worldwide authority’ are similarly meaningless. Other accolades mean far less than they suggest: ‘multi-award-winning scientist’ only requires you to have won two awards of dubious prestige, and an ‘international keynote speaker’ could have given a solitary talk outside his home country.

Our role in the system

Now more than ever, the person on the street plays an important role in combating or amplifying misinformation. A single citizen sharing a flimsy paper or conspiracy theory can help it go viral; even for a careful study, someone might exaggerate its conclusions and this misportrayal spreads. Doctors take the Hippocratic Oath of ‘First do no harm,’ and it’s a useful rule for anyone active on social media, because what we post is potentially contagious.

One guideline is to *pause before sharing*. X now asks users to read an article before reposting it; they found that people open it up 40% more often after seeing the prompt.[14] That’s good practice, but it doesn’t go far enough. Even if we read an article, we may take it at face value. If we don’t have time or expertise to do the checks in Part II, we should be very careful about sharing it, as we might be spreading untruths. Many people’s X profiles say ‘Retweet is not endorsement,’ but that’s a cop-out. Even if it’s not an endorsement of the position – I sometimes share anti-sustainability articles if they’re well founded – it should be an endorsement of the analysis.

You might worry that pausing won’t achieve much. If the reason we share misinformation is it’s just too difficult to figure out whether something is accurate, taking a couple of deep breaths won't help. A study by Gordon Pennycook and co-authors reached a more positive conclusion.[15] They gave 1,015 people a set of news stories, half of which were true and the other half false, and asked them to rate their accuracy. Participants did a good job in separating fact from fiction. The researchers then asked them whether they intended to share the article on social media. Worryingly, their answer depended on whether it agreed with their political beliefs rather than whether they'd deemed it accurate. People have the ability to tell truth from lies, but ignore it when it comes to sharing.

How can we ensure they use this ability? In a separate experiment, the researchers took Twitter (now X) users who frequently reposted links to Breitbart News and InfoWars, two sites widely viewed as untrustworthy. They created new Twitter accounts that followed these users. If they were followed back, they sent a direct message like 'Thanks for following me! Can I ask you a favour? I'm wondering how accurate the above headline is, and I'm doing a survey to find out.' It attached a headline and asked the user to rate its accuracy. The authors didn't care about the actual rating but rather the user's subsequent reposting behaviour. Strikingly, they found a significant decline in the sharing of misinformation. Simply getting someone to think about accuracy implanted it in her mind. Then, when she later thought about sharing a story, even on a different topic, she first asked herself if it was truthful.

That’s something we can learn to do ourselves, without any priming.

If, after the accuracy check, we decide to share something, we should then ensure we don’t climb the Ladder of Misinference. A person might post ‘New study finds proof that exercise improves IQ’ when it only contains evidence or data, or ‘The tragic case of Anh Nguyen proves that this country is hostile to foreigners’ when it’s only a single fact. After the article ‘Boardrooms with more women deliver more on climate’ came out, which we saw in Chapter 3, my LinkedIn feed had posts such as ‘Diverse boards do better period’ and ‘Boards with more women on them deliver more value. Period.’ Quite apart from there being no study, ‘period’ implies the issue has been proven so there’s no room for another opinion – ironically, the opposite of diversity.

A second guideline is to *pause before criticizing*. A paper whose conclusions we don’t like triggers our amygdala, and we’re raring to demolish it. When such research is shared, naysayers will carelessly trot out the phrase ‘Correlation is not causation’ – without reading the article and seeing if the researchers addressed this concern. Often readers will criticize a study by starting with ‘I haven’t read it, but . . .’ That’s like a restaurant review declaring ‘I haven’t actually eaten there, but the food’s lousy.’

Just as academics, authors and companies have incentives other than the truth, so do the rest of us. If our goal is to maximize our likes or reposts, we’ll craft black-

and-white posts that prey on confirmation bias, claim to have found proof and end with ‘period’ to imply the case is closed. Or we’ll chase Likes by cancelling someone who shares an unpopular viewpoint, labelling them Taliban, Flat Earthers or climate-change deniers.

We can do better. Our criterion for writing a post or sharing a study shouldn’t be to gain popularity but to inform. Then, what matters is whether the statement is backed up by fact, the facts are reinforced by large-scale data, and the data addresses alternative explanations and so counts as evidence . . . but the authors stop short of claiming proof.

In a nutshell

- Thinking smarter as individuals involves actively seeking dissenting viewpoints.
    - ◦ For an individual topic, we can search for articles that take the other side.
    - ◦ More generally, we can obtain news from both ends of the political spectrum.
- Peer review lets us stand on the shoulders of giants, who perform the checks in Part II for us. We should be cautious of studies that are not published in top peer-reviewed journals.
    - ◦ Peer review is imperfect. We should avoid putting too

much weight on a single paper and instead learn the *scientific consensus* by reading a *systematic review*.

- Company reports, books and newspaper/magazine articles never try for peer review. We can evaluate:
  - Is it balanced?
  - Does it exaggerate its results or its rigour?
  - Might the authors be biased? What are their incentives to claim the result? Would they have published the opposite finding?
  - How strong are the authors' credentials? What are their research qualifications and track record of publications? Which institutions are they at?
- We all play a role in combating misinformation. This involves refraining from:
  - Sharing a study unless we've vetted it. Reposts should be seen as endorsements of the analysis, if not the position.
  - Climbing the Ladder of Misinference for anything we do share.
  - Dismissing an article without checking if the authors have addressed our concerns.

Many decisions aren't made by individuals; they're made by organizations. An effective organization doesn't just involve members individually reading the best scientific research but also contributing their unique experiences, backgrounds and cultures. The next chapter explains how to harness this diversity of thought, to ensure the whole is more than the sum of its parts.

10

Creating Organizations that Think Smarter

On the morning of Tuesday, 16 October 1962, President John F. Kennedy received alarming news. National Security Advisor McGeorge Bundy informed him that a U-2 spy plane had photographed ballistic missiles being installed in Cuba.

Kennedy was both afraid and enraged. Afraid, because Cuba was just 90 miles south of Florida and these weapons had a range of 1,200 miles. They posed a massive threat to American citizens and could trigger a full-scale war – particularly since tensions between the US and the USSR had been running high for fifteen years. Enraged, because Soviet Premier Nikita Khrushchev had pledged privately and announced publicly that he'd only send defensive weapons to Cuba.

Kennedy had to act fast – but act wisely. Just eighteen months earlier, he'd supported the disastrous Bay of Pigs invasion, a failed attempt to overthrow Cuban dictator

Fidel Castro. A staunch communist, Castro had come to power in 1959 by ousting his predecessor in an armed revolt. Castro’s strong relationship with the USSR, plus his determination to communize other Latin American countries, made him a serious threat. In March 1960, then President Dwight Eisenhower approved a plan, drawn up by the Central Intelligence Agency (CIA), to train up Cuban exiles to invade their homeland and topple Castro.

In February 1961, not long after his inauguration, Kennedy met with his most trusted advisors, and the clear consensus was to go ahead. He was determined to disguise US support, and so chose the remote Bay of Pigs as the invasion point. On 17 April 1961, 1,400 Cuban exiles landed, raring to march on Havana and dethrone the dictator.

But Castro’s intelligence had got wind of the plot, and he was prepared. Cuban planes quickly took to the skies and 20,000 Cuban troops swarmed on the land, crushing the invasion within three days. More than 100 exiles were killed and 1,200 were captured; they were released two years later in return for $53 million worth of food and medicine.

The Bay of Pigs disaster had a lasting impact on Kennedy. Stung by the humiliation, he analysed its origins. One was the complacency of the CIA and the Joint Chiefs of Staff (JCS, the most senior military leaders), who’d assured him the invasion would succeed: the backward Cuban army would be no match for the US-trained and -equipped exiles.

The President blamed their hubris – ‘Those sons of bitches with all the fruit salad just sat there nodding, saying it would work’ – and repeatedly told his wife, ‘Oh my God, the bunch of advisors that we inherited!’ But he saved the greatest blame for himself. The advisors were just that – merely advisors – and the ultimate decision was his. He’d been too trusting, too deferential; and not only did he fail to challenge them himself, he also failed to create an environment where others felt comfortable doing so.

Yale psychologist Irving Janis coined the term ‘groupthink’ to describe what led to the Bay of Pigs calamity.[1] This occurs when members desire to fit in and be accepted by a group. As a result, they’re reluctant to rock the boat and instead support whatever decision they think the team wants. Since the attack had already been approved by Eisenhower, it was the default position that the Kennedy administration anchored on, and so they ‘uncritically accepted’ it, according to Janis. At each meeting, the President allowed the CIA and JCS to dominate the discussion; whenever anyone expressed a concern, he let them shoot it down immediately rather than giving space for others to reinforce the doubt. As a result, dissenters ended up self-censoring.

Kennedy was determined not to repeat his blunder. The very day he heard about the Soviet missiles he assembled a group of fourteen advisors in a body he later named EXCOMM, the Executive Committee of the National Security Council (NSC). It’s easier to disagree with a dozen than fifty, so this smaller team would reduce the risk of groupthink. The President chose the members to ensure a diversity of viewpoints. Even though it was officially the Executive Committee of the NSC, four of the thirteen permanent members were from outside the NSC. It also included a dozen rotating advisors from other federal agencies.

The military chiefs, such as JCS Chairman General Maxwell Taylor, were eager to launch an air strike on the missile sites, followed by a full-scale invasion. EXCOMM ended up choosing a far less aggressive course – a blockade of Cuba to stop new missiles arriving and an ultimatum to remove the existing ones. Khrushchev agreed to the withdrawal in return for Kennedy publicly promising not to invade Cuba again. The following year, the two leaders established a direct telephone hotline to enable immediate contact if tensions rose again.

How did Kennedy avert a nuclear battle that could have escalated into World War Three, when just eighteen months earlier he'd faced global humiliation? Kennedy secretly recorded EXCOMM's deliberations, which have since been declassified and released; social scientists have pored over the transcripts for pointers on how to avoid groupthink. We'll get into the actual conversations shortly, but we'll first focus on a lessstudied element – the composition of the committee, and in particular its cognitive diversity.

The power of different points of view

Cognitive diversity refers to a variety of backgrounds, experiences, beliefs, ways of interpreting information and approaches to solving problems. Most diversity efforts focus on *demographic diversity*, which typically prioritizes gender and ethnicity.[*] They’re related to cognitive diversity – men think differently from women, and different cultures have different values. But often ignored is age. Generations who’ve lived through recessions and financial crises are more cautious than those who’ve only experienced sunshine.[2] Younger cohorts tend to be more tech-savvy, but they may overestimate the power of technology; they also have greater concern for environmental and social issues.[3]

In addition, there are many aspects of cognitive diversity that aren’t captured by demographic characteristics such as gender, race and age. An old white male won’t tick any diversity boxes, but his expertise might be in marketing when his colleagues are accountants, or he might be from a less affluent background and so be better able to understand the concerns of shop-floor employees.

Almost inevitably, given it was 1962, EXCOMM consisted exclusively of white men, which is normally a disaster for groupthink. However, EXCOMM enjoyed an unusual amount of cognitive diversity, and one source was JFK himself. The military leaders naturally saw armed conflict as the solution to most disputes. US Navy chief Arleigh Burke came close to delivering a speech recommending the

US bomb ‘the Soviet Union from hell to breakfast’. Air Force general Thomas Power also had no time for shades of grey: ‘The whole idea is to *kill* the bastards . . . At the end of the war, if there are two Americans and one Russian, we win.’ They applied this view to the Cuban missile situation. On the fourth morning of the crisis, they unanimously recommended the air strike, with Air Force Chief of Staff Curtis LeMay making clear to Kennedy: ‘We don’t have any choice except direct military action.’

Had Eisenhower still been President, that action might well have been taken. Kennedy’s predecessor served as General of the Army during the Second World War and was only the fifth American ever to hold the rank of five-star general. The army chiefs held him in the highest esteem, and he in turn trusted them. The contrast with Kennedy couldn’t have been greater. Inaugurated at the age of 43, Kennedy just didn’t have the same gravitas as Eisenhower, who stepped down at 70. And it wasn’t just age, but military stripes. Lyman Lemnitzer, who’d been succeeded as JCS Chairman by Maxwell Taylor two weeks earlier, said of Kennedy: ‘Here was a president who had no military experience at all, sort of a patrol-boat skipper in World War II.’[4]

But what was a major shortcoming to Lemnitzer was a major gift to humanity, as Kennedy’s background meant that he saw the downsides of using hard power as well as its strengths. When the JCS suggested bombing the Cuban missile sites, Kennedy counter-proposed a blockade, leading LeMay to scoff, ‘This is almost as bad as the appeasement at Munich.’ Kennedy’s advisors also had nuanced views of military action. Defense Secretary Robert McNamara, who stopped Burke giving the speech about bombing the Soviet Union, once opposed demands for additional air forces. This prompted LeMay to scorn, ‘Would things be much worse if Khrushchev were Secretary of Defense?’ Secretary of State Dean Rusk also preferred diplomacy over war. In July 1961, when the JCS proposed a Pearl Harbor-style attack on Russia, Kennedy left the meeting remarking to Rusk, ‘And we call ourselves the human race?’

The idea that cognitive diversity helped avoid World War Three is a compelling story, but it’s only a story. There could be many other reasons why EXCOMM reached the correct decision. Academic research can more precisely pinpoint how cognitive diversity affects group effectiveness. Ishani Aggarwal and co-authors took 337 volunteers and measured their cognitive style – the way they process and use information – using an established method called the Object–Spatial Imagery and Verbal Questionnaire.[5] They then divided the subjects randomly into ninety-eight teams, and tested each team’s ‘collective intelligence’ using tasks from the McGrath Task Circumplex, another accepted framework.[†]

The researchers found that teams with a greater diversity of cognitive styles enjoyed higher collective intelligence – but only up to a point. It’s not black and white; there's moderation. Too much diversity can backfire because team members 'speak different languages' and have a hard time understanding each other.

Picking your fantasy team

How do we build cognitively diverse groups? We can't give every employee the Object–Spatial Imagery and Verbal Questionnaire before deciding who to put on a committee, so most companies need to use simpler measures. A first step is to ensure gender and ethnic diversity, but this doesn't go far enough. As discussed earlier, diversity in age, career paths and socioeconomic backgrounds helps ensure a variety of lived experiences.

Another useful indicator is *social diversity* – belonging to different social groups. Slava Fos, Elisabeth Kempf and Margarita Tsoutsoura identified the political affiliation of top US executives based on donations and voter records.[6] They found that a departure of a politically misaligned executive, such as a Democrat quitting from a mainly Republican-led company, costs the average firm $200 million – potentially due to the increased groupthink.

A different study investigated the channels through which social diversity may matter. They asked 186 people whether they identified as a Democrat or a Republican and then had them read a murder mystery, identify who they thought had committed the crime, and prepare for a meeting with another participant.[7] They were told their partner disagreed and that they'd need to come to a consensus. The first step was to write a statement explaining their view, which the counterparty would read before the discussion. Half the subjects were told their partner belonged to the same political party, the other half to the opposition.

The researchers found that Democrats prepared better for the meeting, as measured by a more comprehensive essay, when they were contradicted by a Republican rather than a fellow Democrat; for Republicans, it was the same (but with the parties switched). These results suggest that social diversity prompts us to work harder to address disagreement. In contrast, if the person is in the same social circle as us, we think we can convince them using only charisma.

While that study looks at preparation, a separate paper co-authored by Katherine Phillips explores outcomes.[8] The researchers took a group of fraternity brothers[‡] and asked them to solve a murder mystery. Five minutes in, they added a newcomer to the group. When he was from outside the house, the group was more likely to correctly identify the perpetrator than when he was a fellow resident.

Interestingly, this superior performance wasn't because the newcomer provided fresh ideas of his own. Instead, it was because his entry changed the dynamics between the existing members. During the initial five-minute discussion, the brothers typically had different views on who the culprit was. When a newcomer from a different fraternity arrived and shared his opinion, members who agreed with him found themselves in an awkward situation – they concurred more with an outsider than their own brothers. This tension made them more willing to try to understand their brothers' opposing views. In contrast, when the newcomer was from the same fraternity, they saw no conflict and kept arguing against their brothers. The results reinforced the findings of Phillips's earlier study: social diversity leads people to take dissenting opinions more seriously.

From diversity to inclusion

Diversity is only a starting point. Many companies take an 'add diversity and stir' approach, where they think it's enough to recruit a mix of people and then sit back and wait for performance to skyrocket. But what matters is diversity *and inclusion* – creating the conditions for diverse colleagues to share their different viewpoints. Without inclusion, an organization will never fully realize the benefits of its diversity. Indeed, we've seen how gender diversity alone doesn't improve performance. This could be because demographic diversity might not tally with cognitive diversity, or because it fails to capture inclusion.

I wanted to move beyond prior studies that focus only on demographic diversity and study the importance of inclusion. The Best Companies to Work for list, which I used for my employee satisfaction research, is compiled after surveying employees at each company on 58 different issues. Only the list itself is publicly disclosed, but Caroline Flammer, Simon Glossner and I obtained confidential access to the individual survey responses. 13 out of the 58 questions are related to diversity, equity and inclusion (DEI), such as ‘This is a psychologically and emotionally healthy place to work,’ ‘I can be myself around here’ and ‘Managers avoid playing favourites.’ (Others explore employee satisfaction in general, such as ‘I am given the resources and equipment to do my job.’)

Caroline, Simon and I found that a company’s score across these 13 DEI questions is unrelated to gender or ethnic diversity in either the boardroom, the CEO position or the wider workforce. Indeed, we uncovered several companies which ticked the box for having high demographic diversity but never improved DEI. This distinction matters: we found that DEI is associated with higher future performance, but demographic diversity is not.[9]

Kennedy, McNamara and Rusk were all present at the Bay of Pigs deliberations, but they failed to prevent the blunder, perhaps because cognitive diversity alone isn’t enough. In the Cuban Missile Crisis, the President and EXCOMM took several steps to ensure that cognitive diversity was accompanied by inclusion. Some members were primed to launch an air strike, but before rushing to debate its pros and cons, EXCOMM first listed the entire range of options – similar to considering alternative hypotheses to your pet theory. The choices were: the air strike, the blockade, doing nothing, pressuring the USSR to remove the missiles, asking Castro to split from the Soviets or be invaded, and a full-scale invasion of Cuba.[10]

After discussing these six, EXCOMM then whittled them down to the air strike and the blockade.[§] They then divided themselves into two groups, each tasked with justifying one of the finalists. As Attorney General Robert Kennedy describes in *Thirteen Days: A Memoir of the Cuban Missile Crisis*, each team wrote up a full paper which began 'with an outline of the President's speech to the nation and the whole course of action thereafter'. The two plans were presented to Kennedy, who chose the blockade.

A second key step was that Kennedy recused himself from many of the meetings, so that the members didn't feel pressured to support whatever action they thought he favoured. Not only did Kennedy not chair the meetings, but there was no chair at all. In Robert Kennedy's words, 'During all these deliberations, we all spoke as equals . . . the conversations were completely uninhibited and unrestricted. Everyone had an equal opportunity to express himself and to be heard directly. It was a tremendously advantageous procedure that does not frequently occur within the executive branch of the government, where rank is often so important.'

Many features of the EXCOMM deliberation process can be applied to any organization. Considering all available responses rather than rushing to discuss just one is the practice of *brainstorming*. Sometimes, brainstorming is undertaken in response to a known problem, such as the Cuban Missile Crisis or the entry of a new competitor. However, brainstorming can also be done without a fire to be fought – some of the best innovations are driven not by problem-solving but problemfinding: discovering a solution to a problem that no one was actively trying to fix. Nobody asked Henry Ford to manufacture a car; he's often credited with the quote 'If I had asked people what they wanted, they would have said faster horses.' Relatedly, *blue-sky thinking* is brainstorming with no limits – sharing ideas that don't need to be grounded in reality. One person may come up with an idea that's impossible for budgetary or technological reasons, but it may inspire another idea that sparks yet another. Finally, you land on one that hits the sweet spot of being both innovative and feasible.

One problem, particularly with blue-sky thinking, is that group members may be unwilling to share bold ideas for fear of looking foolish. It's all well and good Robert Kennedy claiming that everyone spoke as equals, but EXCOMM contained America's most senior decision-makers. They'd risen to the top, so their reputations could shake off making a suggestion that gets shot down. However, juniors in a company may legitimately feel less secure. A simple solution is to *make the suggestions anonymous*, either writing them down on cards to be read by a facilitator or submitting them electronically.

As with most things, anonymity isn’t always preferable – it’s not black and white. You might wish to put more weight on seniors’ views due to their greater experience. But what matters is the logic behind their suggestion, not their grey hair. The *Delphi method*, named after the Oracle of Delphi renowned for her prophecies, makes two improvements to anonymous brainstorming when it’s used for estimation (such as forecasting the sales of a new product or valuing a company). One is that group members give not only their prediction but the rationale behind it. Second, after seeing the anonymous forecasts and justifications from the first round, colleagues make a revised estimate. Those with limited expertise might move towards a more informed prediction, while others will stick to their guns. This process is repeated many times and stops either after a given number of rounds or once consensus is reached.

While Kennedy recused himself from the EXCOMM meetings, leaders might want to be in the room and hear the discussion. If so, they should *allow juniors to speak first*, so they don’t anchor on their bosses’ views. Yet this alone may not be enough. Beforehand, they may have got wind of what the chiefs think about the agenda items, so they say what the seniors want to hear. Amazon thus practises the *silent start* : it releases the pre-reading only at the beginning of the meeting, and everyone spends half an hour reading it quietly.[11] That way, juniors won’t know their superiors’ views, so what they share are genuinely their own opinions.

Another way to ensure that all opinions are heard is to *hold a vote*. This may seem obvious, but it’s often overlooked. For six years I served on an executive committee that made decisions by consensus – by taking the temperature of the room. When I suggested we cast votes, the Chair refused, claiming they were too formal and would jeopardize the friendly culture of the committee. But the temperature of the room is driven by the most heated voices, so the ‘silent majority’ is ignored.

At one meeting, nearly everyone supported an action. When several months had passed and it hadn’t been taken, I asked why; the Chair claimed that the views were mixed. In fact, every female director had been in favour, plus several non-vociferous males. Yet because one forceful man was opposed, the Chair thought the opinion was balanced. To an outsider, the Chair appeared to be a proponent of diversity, given the committee’s gender balance, but this was irrelevant without inclusion. Holding votes ensures that everyone’s opinion is heard; trying to maintain a friendly culture creates an old boys’ club.

The vote should also be held in a way that creates room to dissent. Consider this common situation: a Chair needs majority approval to take a decision, such as appointing a new member. It’s between meetings, so he sends an email with the nomination and asks for sign-off. A few minutes later, someone hits ‘Reply All’ and falls over herself to stress what an amazing choice the Chair has made. Then a second person Replies All agreeing, and the dominoes quickly fall.

You read the proposal carefully, have significant concerns and are poised to vote No. But because you've joined the email trail late, you see that five colleagues have already said Yes. You might feel pressured to vote Yes as well because it's uncomfortable to go against the grain. Strikingly, a landmark economic theory shows that *even if you have no such concerns* – perhaps because there's no stigma to expressing a different view – and your only goal is to help the group reach the right decision, you might still vote Yes.[11] You reason that 'If five colleagues think it's a good idea, I must be making a mistake.' This is known as an 'informational cascade' – a series of votes in the same direction causes you to suppress your own information and follow the herd. The group never benefits from your insights.

The remedy is simple: the Chair asks members to email him their votes privately so that they're based on their own opinions rather than what others selected. If he wishes to hold a group discussion before the vote, he can ask everyone to email their thoughts and then release them all simultaneously so the discussion isn't swayed by whoever happens to hit Send first.

Processing power

Brainstorming and the Delphi method help ensure a diversity of perspectives but require a big time commitment. Even more central to inclusion is getting ‘micro-processes’ right. These are small changes with large impacts – in particular, they help ensure that inclusion is embedded in an organization’s DNA rather than just confined to sessions with marker pens and flip charts.

One practice is to *remove default decisions* so that people don’t anchor on them. I serve on the Responsible Investment Advisory Committee for Royal London Asset Management’s eight sustainable funds. One of our tasks is to give an outside perspective on whether a particular company counts as sustainable, in which case the funds are allowed to invest in it. The internal team used to write a report starting with their assessment and then explaining the arguments for and against their position. That’s consistent with what everyone tells you about good writing – start with the punchline; don’t leave it to the end like a mystery novel.

But this structure meant that we anchored on the team’s view – it was the default. If we knew they’d recommended Exclude, we’d read the report unintentionally overweighting the negatives and downplaying the positives. Now the reports simply lay out the arguments and leave the recommendation to the end. By that time, we’ve already formed our own opinion and can independently decide whether we agree or disagree.

A second micro-process is to *reduce hierarchies*. Airforce pilots used to be called ‘Sir’, which meant that crew members viewed themselves as junior and unable to challenge their authority. Social scientists Joseph Soeters and Peter Boer analysed the air forces of fourteen countries and found that accidents were higher in countries with a greater ‘power distance’ – a measure of how much its citizens accept power differences (it’s high in Mexico and China and low in Denmark and New Zealand).[12] An obsession with rank can cost lives.

Reducing hierarchies can be as simple as lessening title distinctions. In most investment banks, there are multiple grades, such as analyst, associate, vice-president, executive director and managing director. As soon as you receive an email from someone, you look up their rank in the internal directory. Their title determines how much weight you give to a suggestion and how much time you spend on a request. You then value comments not on their quality but on who they came from; you drop everything to indulge a senior person’s curiosity even if his question isn’t business critical.

In 2022, investment bank UBS announced that it would remove titles from its internal directory. Earlier that year, UBS had already scrapped ranks above managing director, such as ‘group managing director’ and ‘divisional vice-chairperson’ – once you’ve made MD, you’re in the top tier, so there’s no need for further distinctions. In my fledgling banking career, a vice-president and I once spent twenty minutes debating whether the head of mergers and acquisitions was more senior than the head of corporate finance, because this determined who'd go first in the 'To' box for an email we wanted to send. The importance of hierarchies had permeated the bank's culture, meaning that even putting someone second in a mass email might be seen as a slight. (We eventually solved the problem by putting one in the 'To' box and another in the 'Cc' box, so that each could bask in the glory of being first.)

A third micro-process is to *tolerate failure*. If mistakes lead to chastisement or a negative stigma, people won't pursue the most daring innovations or even propose them in a brainstorming session. Many companies focus on preventing errors of commission (doing something that fails), but an obsession with ironing out mistakes often leads to errors of omission (not trying something new).

Some companies go further than tolerating mistakes – they actively reward them. They give accolades for ideas that ultimately failed but provided valuable learnings, and hold 'failure parties' to celebrate the takeaways. As Scott Cook, co-founder of software company Intuit, explains: 'Every failure teaches something important that can be the seed for the next great idea.' Pixar has a failure gallery displaying characters, scenes and gags that never made it to the final movie, demonstrating their belief that failures can be a work of art, like a blooper reel on a film. (Beyond failures, Pixar makes it routine to share unfinished animations with colleagues so that they won't be afraid to get feedback on very rough drafts.) Some organizations go public. For ten years, the non-profit Engineers Without Borders released a Failure Report detailing that year's flops – the initial intentions, what happened and the lessons learned.

A final micro-process is to *ask those with strong opinions to articulate them in detail*. Having to explain something precisely can make people realize they don't know it as well as they thought, opening them up to different views. Yale psychologists Leonid Rozenblit and Frank Keil demonstrated this with a study.[13] They took topics such as how a toilet flush operates, how piano keys make sounds and how a helicopter flies and asked students to rate their knowledge. Most awarded themselves a high score. Then they had to write a step-by-step explanation of how these actually work and, afterwards, re-rate their knowledge. Humbled by their inadequate explanations, they lowered their marks.

You might think that knowing how a toilet operates is different from deciding whether to bomb Cuba: one's an objective description of current reality, the other's a subjective opinion on a future action. Philip Fernbach and co-authors conducted a similar study replacing household items with public policy questions, such as whether there should be a national flat tax or performance-based pay for teachers.[14] The other change was asking participants to rate not only their understanding of the issue but the strength of their stance. Being forced to explain their position reduced not only subjects' estimation of their own understanding, as in Rozenblit and Keil's original study, but also their extremism – making them more willing to listen to alternative opinions.

Amazon applies these findings in their 'silent start'. The prereading is a full-prose memo rather than a PowerPoint deck. With PowerPoint, you can write down a few bullet points of jargon to create the illusion of knowledge – perhaps copy-and-pasting them from Google even if you don't understand them. A memo requires you to construct an argument and explain your reasoning, deterring you from using a buzzword or making a point unless you can fully back it up.

The devil's advocate

My heart was beating out of my chest and I was breaking into a cold sweat. You'd expect nerves from a Ph.D. student presenting at a conference full of professors, but I'd just finished my talk. I'd practised that presentation dozens of times and never quite nailed it. Yet game day had come and I'd hit every note, so why the jitters?

As I took my seat, Patrick Bolton stood up. He was one of the world's most respected researchers in corporate finance theory, the topic I'd just spoken about, and would later become President of the American Finance Association, the most prestigious position in my profession. Patrick wasn't presenting one of his own papers but instead was assigned the role of ‘discussant’ – to read my paper beforehand and give an independent view.

The discussion started well enough, with Patrick calling my idea ‘intuitively plausible’, but then he said a key ingredient in my study ‘makes no sense’. I felt like I’d been punched in the stomach. I’d been thrilled to be invited to this conference, where I was the only student in a room full of professors. But now they’d be going back home and telling their colleagues about this young upstart who gatecrashed an event for seasoned faculty and presented nonsense.

Patrick then moved on to discussing the next paper in the session, but I was so upset I tuned it out. This research was by a senior professor, so Patrick would surely be full of praise, making mine seem even more flawed by comparison. I was suddenly shaken out of my slumber when Patrick said that this study might have an endogeneity problem. Did I hear that right? As we saw in Chapter 6, an endogenous input knocks you down a rung from evidence to data. I looked up to the slides and, true enough, ‘endogeneity problem’ was up there, as clear as day.

That was the first morning of the week-long conference. Virtually every other discussion followed a similar tone to Patrick’s. It started off commending the question the researchers were exploring, but then explained why they hadn’t yet fully nailed the answer due to alternative explanations or other quibbles. Over that week, I realized that constructive criticism is simply part of the academic process. The whole point of presenting at a conference is that you can only take an idea so far by yourself. There’s no stigma in receiving negative comments – they’re simply expected. If a discussant were ever entirely positive, it would have so little credibility that the audience would think you had incriminating photos of him.

The value of this practice applies far beyond academia. While Part II highlighted the power of the scientific method, here we stress the value of the *scientific culture* – an environment where people put out bold and innovative ideas, actively seek dissenting opinions and revise their proposals to address the criticisms – which is valuable to any organization. If it’s part of the fabric for plans to be critiqued, then there’s no shame in receiving pushback. Nor is there fear in raising concerns – doing so helps colleagues refine their ideas, rather than stabbing them in the back. Highlighting flaws isn’t unkind; instead, one of the most unkind things you can do is to notice a problem and not point it out.

The non-academic equivalent of a discussant is a *devil’s advocate*, who’s appointed to highlight the blind spots in a proposal.[#] Sometimes, an entire group is tasked with this job, known as a ‘red team’.[**] This was practised by EXCOMM: after the two sides wrote their initial papers, they exchanged them and each gave feedback on the other’s proposal. The initial group then revised their plan to take into account the concerns.

Sometimes you don’t need to appoint a red team; the culture is such that one naturally emerges. When he ran General Motors, Alfred Sloan closed a meeting by asking ‘I take it we are all in complete agreement on the decision here?’ Everyone nodded. Sloan continued, ‘Then, I propose we postpone further discussion of this matter until our next meeting to give ourselves time to develop disagreement and perhaps gain some understanding of what this decision is about.’ He believed that no decision is black and white, and if no one raised any concerns, this wasn’t because there weren’t any but because he hadn’t yet given his colleagues time to think of them.

Devil’s advocates can even be automated. Singaporean bank DBS brings a Wreckoon, a racoon-themed mascot wielding a hammer, into every major meeting. At random times, a Power-Point slide appears with the Wreckoon, accompanied by a question such as ‘What have we missed out?’, ‘What is our riskiest assumption?’, ‘What could go wrong?’ and ‘Where is the data?’ This prompts leaders to pause and give airtime to dissenting views.

Knowing that criticism will come your way drives you to make your idea as strong as possible beforehand. Researchers will do all they can to pick holes in their own paper before sending it to a discussant. This practice is known as a *premortem*. In a post-mortem, a decision has flopped and you try to figure out why. In a pre-mortem, you imagine that a failure has occurred and think about all the possible causes – for example, the CEO left and the strategy relied too much on her vision.

A final feature of a scientific culture is the *value given to dissenting voices*. You might think it strange that people ever agree to be a discussant – you fly halfway around the world to be the bad guy in the room – but the profession greatly respects members who give tough but constructive evaluations. Doing so boosts their reputation, and many conferences give ‘best discussant’ awards.

Some companies aim to foster such a culture. X, Google’s moonshot factory, gives a bonus to any employee who finds a fatal flaw that leads to their own team’s project being killed. This in turn inspires X’s engineers to be yet more daring – if they propose a crazy idea that has a fundamental defect, they’re confident that a colleague will notice it and scrap the innovation before it costs the company millions of dollars. The better a car’s brakes, the more you can push on the accelerator.

This recognition should extend beyond outright dissent to simply sharing a different perspective. In my time at Morgan Stanley, the half-yearly evaluation summaries were a six-by-two table. The left column contained three strengths and three development areas; the right was reserved for unusual remarks and nearly always blank. I only ever received one ‘right column’ comment, and it came in my first evaluation: ‘Notably for a first-year analyst, Alex has the confidence to speak up and express his own views, which is to be encouraged.’ I was stunned.

Everyone says that the ideal junior is someone who does what they're told, no questions asked. This single sentence corrected my misperception – it made it clear that the bank valued our perspectives, no matter how inexperienced we were.

Beyond formal evaluations, at the end of a meeting the chair could privately thank members who raised a contrary opinion and acknowledge the courage it may have taken to do so. This is particularly valuable if the concerns didn't change the decision, so the colleagues might think they weren't valued and the effort was for nothing – or, worse still, was an annoyance that prolonged the meeting unnecessarily.

Yet not every company values dissent. In May 2022, Stuart Kirk, HSBC's head of responsible investment, gave a speech arguing that investors needn't worry about climate change. This talk led to instant outrage, but the content was more nuanced than the headlines suggested. He pointed out that, even if the planet becomes warmer, we can invest in adapting to higher temperatures.[††] Nor did he say that climate change isn't a serious threat to *society* but rather that *investors* don't bear the risks as their horizons are too short-term. HSBC suspended him, even though they'd previously signed off on the content of his talk, and the furore led to him resigning shortly afterwards.

Kirk's delivery was sometimes sardonic, with the mostquoted line being 'Who cares if Miami is six metres underwater in 100 years? Amsterdam has been six metres underwater for ages, and that's a really nice place. We will cope with it.' However, controlling our emotions about the tone and focusing instead on the content, the speech did an important service by providing a contrasting opinion – that we're focusing almost exclusively on climate-change mitigation and not enough on adaptation, and that investors won't worry enough about climate change until regulators make them pay the price through carbon taxes. Suspending someone for expressing a dissenting view, even on a topic we might feel strongly about, is a deterrent to diverse thinking.

In a nutshell

- Creating organizations that think smarter involves overcoming *groupthink* : the desire to conform.
- Overcoming groupthink involves recruiting colleagues with cognitive, not just demographic diversity and taking active steps to harness their collective wisdom.
- Formal procedures include:
  - Brainstorming sessions for problem-solving: considering all available options before focusing on particular ones.
  - Blue-sky thinking for problem-finding, unconstrained by budgetary or technical feasibility.
  - 'Silent starts', where the agenda and papers are only

released at the start of the meeting. When the discussion begins, juniors speak first.

    - Anonymous votes on important decisions rather than just taking the temperature of the room.
    - The Delphi method for forecasting, where forecasts are accompanied by rationales.

- Informal micro-processes ensure that everyone is listened to, not just heard, and is willing to share bold or contrarian ideas. Examples include:
    - Removing defaults.
    - Flattening hierarchies and playing down job titles.
    - Celebrating constructive failure.
    - Asking people with strong opinions to explain their position in detail.

- A *scientific culture* proposes innovative ideas and actively seeks criticism. This involves:
    - Red teaming: appointing a group to play devil's advocate.
    - Rewarding dissent and respecting dissenters, so that devil's advocates emerge.
    - Pre-mortems: imagining an idea has failed and asking why.

Just as organizations are more than a group of individuals, societies are more than a set of organizations. In our final chapter, we'll explore how to create societies that think smarter.

11

Creating Societies that Think Smarter

‘Education, education, education.’ With these words, UK Labour party leader Tony Blair set out his party’s priorities during the 1997 General Election campaign. Could the same approach hold the key to building smarter-thinking societies? With knowledge at our fingertips, surely we’ll be able to defend ourselves against the arrows of misinformation?

As you’re probably used to by now, the answer isn’t quite this simple. In Chapter 1 we saw that education in fact makes biases worse by equipping us to engage in motivated reasoning. Yet to conclude that no form of education can help would be black-and-white thinking. General schooling and standard numeracy aren’t effective, but teaching targeted at misinformation and misinference could still do the trick. So what types of education might work? As in medicine, diagnosis precedes treatment – to understand how best to solve a problem, we must first identify its causes.

We previously encountered the Lord, Ross and Lepper study where people lapped up papers that supported their position on the death penalty but rejected those that contradicted it. Lord and Lepper teamed up with Elizabeth Preston to investigate what causes such biased interpretation.[1] One explanation was *awareness* – the students had no idea they were being partisan. The second was *cognitive* – they wanted to get it right but didn't know how to interpret the data correctly.

This new research ran the same experiment as before but asked some students to be 'as *objective* and *unbiased* as possible', to increase awareness of their biases. Others were handed a cognitive tip: to consider how they'd react if the study had found the opposite result.

What does this mean? Normally, if a death-penalty opponent saw that murders were higher in states with capital punishment, she'd instantly interpret this as supporting her viewpoint. 'Consider the opposite' means asking herself how she'd respond had the research found *lower* murder rates in the death-penalty states. She'd appeal to alternative explanations – perhaps these states have stronger economic conditions, and that's why crime is lower, not the death penalty. Now that she's alert to rival theories, she realizes that they might also plague the study's true results and so she no longer takes them at face value. The research actually found that homicides are higher in death-penalty states, but she's now aware this could be because their economies are weak.

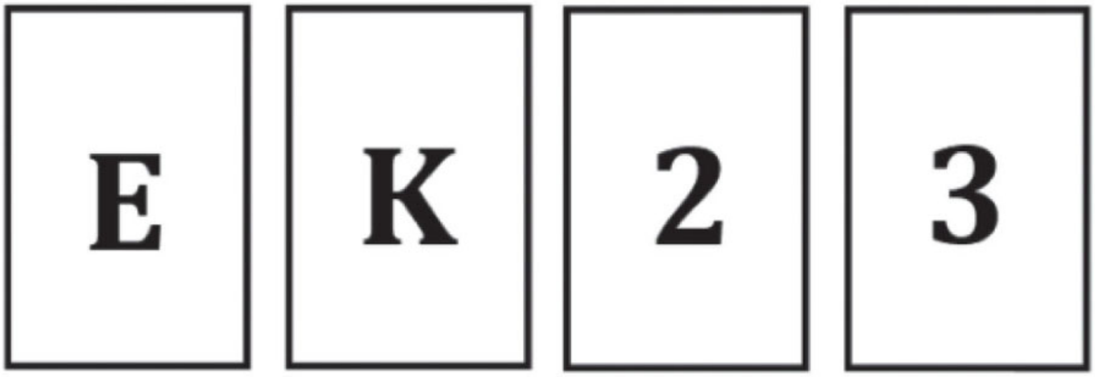
Charles, Mark and Elizabeth found that the 'be unbiased' instructions had no effect – students continued to judge research as more convincing if it supported their views. But 'consider the opposite' cured this biased interpretation: people's assessment of a study no longer depended on whether they liked its findings. The researchers then conducted a separate experiment to demonstrate that 'consider the opposite' also helps overcome biased search.

'Consider the opposite' echoes other bias-beating tactics we've previously encountered. Peter Wason's 2–4–6 problem showed how we should try to disprove our theory of successive even numbers rather than support it. Chapter 9 suggested that we ask ourselves if we'd believe a study even if it found the opposite result, and whether the authors would have still published it. In all these cases, looking through the other end of the telescope helps us see more clearly.

The criticality of critical thinking

Charles, Mark and Elizabeth's results are both striking and encouraging. They suggest that misinformation can be overcome, but it's not as simple as just making people aware of their biases – they're too ingrained for people to respond to that, which is why this book didn't stop at Part I. Instead, the solution is to teach specific skills that develop critical thinking, and it starts in the classroom. *Consider the opposite* should be in every school curriculum: it trains kids to be sceptical of flimsy studies, to embrace different viewpoints and to challenge their own theories. This technique can be coached using logic problems, like how we develop problem-solving using riddles such as the fox, chicken and grain river-crossing problem.[*]

Peter Wason, who invented the 2–4–6 brainteaser, developed another famous puzzle to highlight the power of considering the opposite. You're given four double-sided cards, which you're told have a letter on one side and a number on the other. You see the following:



Which two cards do you need to turn over to test the following rule? 'If a card has a vowel on one side, it has an even number on the other.'

The first card is easy – E – an odd number on the reverse would disprove the rule. Most people's second card is 2. The rule mentions a vowel and an even number, so it's instinctive for E and 2 to be our two cards. We’re hoping to find a vowel on the back of 2, but this would only be *consistent* with the rule – it wouldn’t *support* it. If there’s also a vowel on the back of 3, the rule would be violated. Thus, the second card we should turn over is actually 3, to try to refute the rule. Uncovering a vowel disproves it, but finding a consonant fails to contradict the rule – and so supports it. Adding games like this to the school curriculum would help set kids up for a lifetime of critical thinking.

A second vital skill to impart is *statistical literacy* – the difference between facts and data, data and evidence, and evidence and proof. We live in a world overflowing with numbers, but children rarely encounter statistics until GCSE-level Maths. Yet statistical literacy can be taught from a young age – interpreting data is often a simple logic problem that doesn’t require any mathematical ability; this book hasn’t contained a single equation. Understanding that correlation doesn’t imply causation is as simple as being aware of alternative explanations, just as kids’ whodunit brainteasers have multiple potential culprits. Psychologists Geoffrey Fong, David Krantz and Richard Nisbett showed that teaching elementary statistics improved people’s judgement across several everyday problems. Encouragingly, the skills learned were general – participants showed just as much improvement in contexts where the principles weren’t taught as in those where they were.[2]

While statistical literacy provides the means to avoid confirmation bias, *curiosity* gives the motivation to do so. A research team led by Dan Kahan measured scientific curiosity using a number of methods.[3] For example, they asked students to choose a news story to read from one of four sources (the magazine *Science*, the sports website ESPN, Yahoo! Finance, and the celebrity newspaper *Daily Dish*); they also timed how long they were willing to watch a science video before stopping. Participants were then asked how much risk they believed global warming poses to society. As expected, students aligned with the Democratic Party perceived higher risk than Republicans. More surprisingly, as scientific *intelligence* (measured by a separate set of questions) rose, Democrats reported greater risk but Republicans less, consistent with knowledge enabling motivated reasoning. However, scientific *curiosity* had a different effect – the risk estimated by both liberals *and* conservatives increased, and there was no polarization.

There are many ways to develop curiosity. Formal steps involve producing children's films and TV programmes about science, arts and the humanities. As with developing cultures in organizations, developing curiosity also involves informal micro-processes – in particular, constantly encouraging kids to ask questions, not just memorize facts. The Montessori education method involves independence and self-guided learning where children have the space to reflect and challenge. The art teacher at my school encouraged us to ‘look and see’, rather than ‘look and recognize’ – instead of glimpsing a building and recognizing it as a house, to see that it’s different from regular houses because the windows are arched, not rectangular.

This isn’t only relevant for the young. We can keep people curious throughout their lives with documentaries, museums, exhibitions and public lectures that make academic subjects more accessible and less mystical. In Chapter 3 I mentioned Gresham College, a UK institution that provides free lectures to the public on topics such as astronomy, divinity and the environment. The audience ranges from schoolchildren to retirees; the lectures are also available on video replay, as podcasts and in transcript form.

Considering the opposite, statistical literacy and curiosity all help us understand information that we receive. But smart thinking also involves being able to form arguments of our own. Education professor David Perkins found that general schooling merely led to more one-sided viewpoints – people came up with more my-side arguments on thorny topics but there was little effect on other-side arguments.[4] A four-week course dedicated to reasoning successfully increased the number of opposing points, consistent with what we’ve learned about the value of targeted education.[5]

But the strongest results came not from classes but from providing real-time prompts – a technique known as *scaffolding*. David gave subjects one of two questions: ‘Would providing more money for public [i.e. state] schools significantly increase teaching and learning?’ or ‘Would a nuclear freeze significantly reduce the probability of world war?’, and asked them to write down as many thoughts as they could. Then, he offered scaffolds such as ‘You’ve only given arguments for one side. Can you think of counterpoints?’, ‘Try to come up with three more reasons for saying “yes”’, and ‘Read the question again. Go through each of your arguments and check that they’re all relevant.’

It’s not obvious that scaffolds can make a difference. The students had already been told to come up with as many arguments as they could, and the prompts didn’t g ive them any extra facts about public education or nuclear freezes. Yet the number of my-side reasons doubled, and the otherside arguments rose seven-fold. Before the scaffolds, 16% of the students’ points were for the opposing position; afterwards, this rose to 45%. Participants were perfectly capable of coming up with balanced arguments; they just needed to be prompted.

These findings are encouraging. They suggest that the quality of an argument isn’t limited by your knowledge of a topic but by your ability to use that knowledge. Schools teach individual subjects such as Chemistry and History, but reasoning is a general skill that helps pupils fully leverage this subjectspecific expertise. While the scaffolds were provided in real time, they were also entirely generic – so students can be taught to scaffold on their own.

It’s not us against them

‘97% of scientists . . . have now put that to rest. They’ve acknowledged the planet is warming and human activity is contributing to it.’

Barack Obama booms these words at the start of a video titled ‘Senator Ted Cruz is a climate change denier’. If that wasn’t convincing enough, ‘97% OF SCIENTISTS AGREE’ takes up the whole screen eighteen seconds in. Senator Cruz is labelled a ‘CLIMATE CHANGE DENIER’ and ‘RADICAL AND DANGEROUS’. The video couldn’t make its point any more powerfully.

Actually, it could. And the reasons why apply not only to this video, but also to similar messages given in other forms. At least six high-profile studies were published between 2004 and 2012 highlighting the scientific consensus on global warming;[6] the 2006 Academy Award-winning documentary *An Inconvenient Truth* paints a similar picture. Yet public understanding of climate change hasn’t improved. In 2003, a Gallup poll found that 61% of people believed that climate change is man-made rather than natural; by 2013 this had dropped to 57%. Of course, this data isn’t conclusive, because we don’t know the counterfactual – what it would have been without those studies and that movie – but there hasn’t been the epiphany we’d have hoped for.

Dan Kahan’s explanation is the *cultural cognition hypothesis*. People respond to a message based not on the evidence behind it but on the cultural identity it signifies.

The problem with the video ridiculing Cruz is that it made climate change an issue not just of science but of politics, suggesting that liberals believe in it and conservatives don’t. A Republican viewer might think he needs to be a sceptic if he wants to call himself a proper Republican, irrespective of what the science says. *An Inconvenient Truth* was also loaded with evidence, yet because it was about Al Gore, it politicized the topic. Climate change then becomes less about ‘What do you believe?’ and more about ‘What group do you belong to?’

The cultural cognition hypothesis means that we can’t think about facts, data and evidence in a vacuum. Data is never just data; we don’t evaluate it only on its quality but on whether it supports ‘them’ or ‘us’. Similarly, we support or oppose conclusions not based on the evidence behind them but on whether they’re the sort of things ‘people like them’ or ‘people like us’ tend to say. As a result, to create more informed, smarter-thinking societies, public messaging needs to disentangle evidence from identity.

One step is to resist the temptation to ridicule opponents, such as laughing at Republicans for ignoring a 97% scientific consensus. While doing so might give a short-term dopamine hit, it makes it much harder for the other side to focus on the evidence.

In July 2022, the Oxford professor who labelled sustainability sceptics Taliban and Flat Earthers wrote a *Forbes* article entitled ‘A tutorial on [sustainable] investing in the oil and gas industry for Mr. Pence and his friends’. If the goal of the article was really to educate sustainability sceptics, it failed at the first hurdle. As well as patronizing the intended audience by suggesting they needed a tutorial, it also politicized the issue, implying that true Republicans should be anti-sustainability. The actual aim might have been not to inform sceptics but to win Likes from sustainability cheerleaders; if so, portraying the other side as 'them' and yours as 'us' was the perfect tactic. Indeed, he received several hundred Likes and many positive comments in the spirit of 'Yeah, you tell 'em! Sock it to 'em!' Yet his article did little to persuade anyone who wasn't already a believer.

A second, and more positive, action is to ensure that important messages are given by people with a different political stance (or no stance at all), such as a conservative highlighting global warming or a doctor explaining what America's Affordable Health Choices Act in fact proposed. In June 2021, Republican Representative John Curtis launched a Conservative Climate Caucus to encourage his party to take climate change seriously. It was backed by nearly a third of the Republicans in the House of Representatives.

A study by the Yale Law School's Cultural Cognition Project investigated mandatory vaccination against the human papillomavirus (HPV).[7] As we'd expect, right-wing students were more likely to oppose vaccination than their left-wing counterparts.[†] Particularly interesting was a second experiment in which, before giving their views, people first read arguments on both sides from experts who’d likely be identified as left or right wing based on their fictional profiles (photos and lists of books written).[‡]

When the expert with the left-wing profile gave the provaccination case and the right-wing guru opposed it, the gulf between the views of right-and left-wing participants widened. If instead the right-wing advocate defended vaccination and the left-wing commentator opposed it, polarization shrank.

A third approach is to shift the focus from problems to solutions. People are more willing to acknowledge the disease if they agree with the cure. Another study co-authored by Kahan found that right-wing participants were more willing to accept that climate change is a serious threat if the remedy is geoengineering – launching solar reflectors, injecting aerosol particulates into the stratosphere and capturing carbon to store it in deep geological formations – rather than regulation. This solution links climate change to cultural meanings of human ingenuity and industrial innovation, which resonate with free marketers and help them view climate action as an opportunity, not just a threat.[8]

Checking the facts . . . the data and the evidence

‘*New York Times* bestseller. *Wall Street Journal* bestseller. *USA Today* bestseller . . . *Forbes* 15 Best Business Books of

2015. *Business Insider* 20 Best Business Books of 2015 . . . World Economic Forum #1 pick for Books on Leadership to Read This Holiday.' That's just six of the eighteen accolades on the Amazon page for Amy Cuddy's *Presence*. They're followed by glowing endorsements from prestigious newspapers and influential authors. With such praise, who wouldn't buy her book?

Presence is based on the flawed research we encountered in Chapter 9. Dana Carney, the lead author of Cuddy's original paper, released a statement acknowledging that 'I do not believe that "power pose effects" are real . . . the evidence against the existence of power poses is undeniable.' Cuddy left her Harvard position in 2017 and TED now publishes a health warning next to her talk, caveating 'NOTE: Some of the findings presented in this talk have been referenced in an ongoing debate among social scientists about robustness and reproducibility.'[§]

Yet *Presence* continues to sell strongly, and at the time of writing is an Amazon 'Editors' Pick' for Best Non-fiction – eight years after the publication of the first paper overturning Cuddy's work.[9] In fact, the debunking had already been out nine months before *Presence* hit the shelves.

Why did so many people buy Cuddy's book when the research had already been discredited? Presumably, it's because they had no idea. The problem is there's no easy way to look up which books are built on rock and which stand on sand. If a journal article is later overturned, the same journal – *Psychological Science*, in Cuddy's case – is often willing to publish the critique. If a paper is retracted, a notice appears on the journal's webpage; the website Retraction Watch compiles a centralized database of retractions. However, the typical Amazon customer is unlikely to visit *Psychological Science* or Retraction Watch, or even know that these websites exist.

There are individual webpages that fact-check books. Researcher Alexey Guzey's website contains an article, 'Matthew Walker's "Why We Sleep" is riddled with scientific and factual errors', describing problems like the cropped bar chart, but many readers won't know it exists, as it's not linked to anything. So even if you've heard about the problems with *Presence* and want to do due diligence before buying *Why We Sleep*, you won't know where to look.

One solution is a centralized fact-checking website for books, similar to Retraction Watch for research. Such a website is urgently needed because most people read books, not academic papers – and because, as we've seen, the range of ways in which books lie to us is vast. They may be based on flawed research by the same author (*Presence*), misportray research by other authors (*Outliers*, *Why We Sleep*), conduct their own flimsy research (*Built to Last*, *The Spirit Level*), or abandon research entirely and over-extrapolate from anecdotes (*Start with Why*) or shoot from the hip (*Getting Things Done*).

Such fact-checking websites already exist in other contexts. Quote Investigator verifies whether quotes were actually made by the people they're attributed to and digs into the context. Full Fact in the UK, PolitiFact in the US and Pagella Politica in Italy verify claims made by politicians, journalists and public institutions. Several media companies have fact-checking arms, such as BBC Verify and Reuters' Fact Check in the UK, and Agence France-Presse's Fact Check and *Le Monde*'s Les Décodeurs in France. Some of these sites are charities, others have paying subscribers, and others still are financed by companies, so there are various options to fund one for books.

These sites make a difference – and not only to our ability to spot dubious statements but also to their being made in the first place. Political scientists Brendan Nyhan and Jason Reifler randomly sent letters to state legislators reminding them that PolitiFact was active in their state and that their reputations would be tarnished if they were caught lying. Recipients were significantly less likely to make questionable statements, compared to other legislators sent a placebo letter, or no letter at all.[10]

While fact-checking websites are useful, readers still have to proactively visit them. Recent innovations bring the horse to water. When Donald Trump tweeted that '*There is NO WAY (ZERO!) that Mail-In Ballots will be anything less than substantially fraudulent* ' in May 2020, Twitter added a fact-checking link at the bottom. It didn't state that the tweet was false but simply offered the option to '*Get the facts about mail-in voting*', which linked to a CNN website that disproved Trump's claim.

Emeric Henry, Ekaterina Zhuravskaya and Sergei Guriev found that these caveats successfully reduce the sharing of fake news.[11] In May 2019, they showed 2,537 French Facebook users two misleading statements about the EU by Marine Le Pen's far-right party Rassemblement National.[‖] A third of the subjects were shown fact-checking information, another third were given the option to access this information, and the control group had no access at all. All were then invited to share the original statements on their Facebook page. Both automatic and voluntary fact-checking reduced sharing by 45% compared to the control group.

These findings are encouraging. There are certainly people whose confirmation bias is so strong that facts won't make a difference, as seen in the Lord, Ross and Lepper study. But those researchers took students who had a strong pre-existing view on capital punishment. Emeric, Ekaterina and Sergei considered the general population – some may have been staunch supporters of Rassemblement National, but enough others were sufficiently open-minded that fact-checking had a large effect.

You might think that a more effective approach than inviting a user to 'Get the facts' would be to label a story as false. Facebook takes this approach, flagging some posts as 'Disputed by 3rd-Party Fact-Checkers'. Yet an experiment co-authored by Gordon Pennycook found a catch.[12] They gave subjects twenty-four headlines, half of which were true. For the test group, six of the twelve false headlines had warnings attached to them. As we'd expect, this reduced how accurately they perceived them, compared to a control group that didn't see tags on any of the twenty-four stories.

What was striking was how the test group rated the six other false headlines that weren't tagged. They viewed them as more accurate than the control group who didn't see any warnings at all. Knowing that Facebook flagged some stories as false encouraged the test group to get lazy and accept untagged headlines as true – the 'implied truth effect'. Since fake news is produced much faster than fact-checkers can flag it, the gains from calling out the falsehoods you do catch might be outweighed by the losses on the ones you don't.

Fortunately, the researchers found a solution. In a separate experiment, they not only tagged some stories as false but also labelled others as true, and this eliminated the implied truth effect. When there are only black marks, no news is seen as good news – but if examiners hand out gold stars as well, no news is indeed no news.

Best-practice guidelines

We've previously discussed how misinformation is difficult to regulate. As a result, establishing best practices is essential for any career that involves creating or disseminating research.[#] The relevant professions already have guidelines, but they don’t get to the heart of the problem. Starting with disseminators, the National Union of Journalists’ code of conduct highlights the importance of checking facts, but we’ve seen how this isn’t enough – even if the facts are accurate, people can make misleading claims based on them. A simple but powerful addition would be that journalists shouldn’t cite a study unless they can link to it. This ensures the study actually exists, and is publicly available so that readers don’t have to take the researchers’ claims, or the journalist’s interpretation, at face value.

Moving to creators, academic bodies such as the Financial Management Association have codes of conduct. However, they typically focus on research integrity and interactions with other academics, not the outside world. As chair of the FMA Ethics Committee, I’ve added new guidelines, such as that researchers shouldn’t pitch a paper to the media unless they’ve made the full version freely available.

The most important codes of conduct may surround the book publication process, since books are a major channel through which people learn about research. In rare instances, books are sufficiently fraudulent that they are withdrawn, like *The Whole Pantry*. In most others, they’re not. While Cuddy’s research was debunked, not every single word in *Presence* was based on it, and reasonable people might argue that there’s sufficient merit to justify its continued sale. If so, the publisher or retailers should feature a ‘health warning’, similar to TED’s caveat on Cuddy’s talk.

One important – but rarely scrutinized – aspect of the publication process is endorsements.** Featuring prominently on a book’s cover, as well as on retailer websites far higher up than independent customer reviews, they can influence our buying decisions more than the actual content. But endorsers can wax lyrical about a book without any accountability. Dozens of luminaries endorsed *Presence*, yet their reputations are untarnished. There are strong incentives to write endorsements – the authors may cover your work favourably in return; having your name on a cover (particularly accompanied by ‘bestselling author of X, Y and Z’) gives you publicity. Some people rubber-stamp literally hundreds, often without looking beyond the blurb. A best practice is never to endorse a book unless you’ve actually read it, and (for serial endorsers) to publish a list of books you’ve endorsed on your website so readers can assess how discerningly you lend your name.

This issue applies to endorsing anything, not just books. Organizations releasing a fresh study, proposing a new law or inventing a novel management practice will get gurus to recommend it. Even if they have little expertise in the issue and haven’t fully scrutinized it, they get exposure from doing so and are almost never held accountable if it turns out to be bunkum. For product endorsements, the problems of both limited expertise and skewed incentives are even more pronounced. A typical Instagram influencer has little knowledge of the science behind the products they’re pushing, yet their reward is far greater than publicity: cold hard cash, and lots of it.

All this means is that we should take most endorsements with a grain of salt. This doesn’t mean that they can never be genuine, but the less knowledgeable the endorser, the stronger their incentives, and the more liberal they are with their praise, the more sceptical we should be.

Civil discourse

My head hadn’t stopped spinning for half an hour, yet I’d barely said a word. I was serving as an expert witness in a damages lawsuit, valuing the losses caused by a breach of contract. The other side had told the law firm that hired me that they wanted to discuss my estimates. This sounded reasonable, so along I went.

We’d barely sat down after the obligatory handshakes when the opposition partner, James, launched into a tirade. It was directed almost entirely at Richard, the head of the law firm I was working for, but I was collateral damage. James told Richard that my analysis was garbage, ignoring or not caring that I was in the room; declared that Richard had no case to begin with; and accused Richard of misleading him on various points. Richard’s tone was less confrontational, but still firm. They continued for thirty minutes, battling it out like wild animals fighting for dominance, while I sat in stunned silence.

After James had expended all his testosterone, I finally got the chance to speak. I reasoned: ‘I understand your goal for this meeting is to hear how I came up with my valuation. That’s my goal too. I recognize that you’re unhappy with many things.’ I then went through each of James’s criticisms, one by one, ending with ‘I’m happy to talk through my rationale for all the above. But first, I wanted to check I have the full list of all your concerns, so that I don’t miss out anything.’

James’s face changed. For the first time in the meeting, he no longer looked like he wanted to kill someone. He sat back in his chair and even started to relax a little. ‘Yes, you’re right, I’m unhappy with all these things.’ I then walked through my explanation for the first point and asked James whether he was satisfied. He conceded: ‘I’m not a finance person so I can’t vouch for this. But yes, on the surface, it does seem to make sense.’ I then moved to the next two concerns, and by the end we’d achieved our goal for the meeting. Richard and I would never have got James to fully agree with us – reasonable people can come up with different assumptions – but at least he no longer thought we were lobbing grenades with a blindfold on. The case was settled shortly afterwards with no further meetings.

Like most people, my default reaction to a dispute is to go on the attack. I call the other person wrong or respond defensively to an innocent question. This was a rare case when I was able to control my amygdala. In any dispute, the best approach involves three steps: to take a deep breath and tame your gut response; to emphasize your common ground with the other side; and to position your arguments as seeking to achieve these shared objectives. The goal should be to come to an understanding, rather than to prove your counterparty wrong.

This reminds us of the Aesop fable where the Wind and the Sun quarrel over who’s stronger. They agree to settle it by trying to strip a traveller of his coat. The Wind sends a powerful blast, which nearly blows the coat clean off, but it causes the traveller to grab on more tightly. The harder he puffs, the tighter the man holds on, and so the Wind fails. It’s now the Sun’s turn, and she begins to shine. The traveller takes off his hat and wipes his face, and eventually becomes so hot he removes his coat. The moral: ‘Gentleness and kind persuasion win where force and bluster fail.’

Most of the remedies in this chapter can only be put into place by those in positions of power. Only the Department of Education, or the headteacher at a private school, can add critical thinking to a curriculum; only a company or philanthropic entrepreneur could finance a fact-checking website for books. But other solutions are within our control. Society isn’t exogenous – something that’s randomly out there and we just have to deal with it – it’s endogenous. As members of society, we play a role in shaping it.

The society that's most relevant for us are the people that we interact with and occasionally come into conflict with – at work, in our neighbourhood or online. Those disputes make us angry, sometimes causing us to unfriend a contact we've known for years or destroy a potential business relationship by flaming someone on LinkedIn. The simple way to avoid such unnecessary arguments is to understand that the arguments are indeed unnecessary – to recognize that both parties often have common goals but just different views on how to get there. Brexiters and Remainers both wanted what was best for the UK, and similarly with Republicans and Democrats in the US. Even if you disagree with their proposed solution, it might contain a grain of sense, and you'd learn something from hearing it – from listening with the intent to understand, not the intent to reply, in Covey's words.

How does the idea of respecting different opinions square with Part II? That section suggested that there's one right way of doing things. If authors quote facts, they must be accurate; if they present data, they should have a representative sample and control group; and if they claim evidence, they need to control for common causes and ideally deploy an instrument or natural experiment.

But this one right way is often difficult to implement in practice – valid instruments and natural experiments are hard to find, and even if you have evidence, it's not proof.

As a result, there are very few topics we can be dogmatic about. There's only one right answer for whether carbon or oxygen has a higher atomic mass, because this can be proven. However, you can't prove beyond doubt whether diversity, sustainability or adopted CEOs improve company performance. A single paper is rarely the last word; two people can look at the same body of evidence and reach different conclusions, just as two unbiased jurors in a trial can hear the same testimony and support different verdicts. As a result, contrasting opinions needn't be incorrect; they're like looking at a landscape from a different perspective.

Even though this book's focus has been on how to climb the ladder from statements to facts to data to evidence, my final message is to recognize that even evidence is not the alpha and the omega. If we claim we're taking an action or holding a view exclusively due to the evidence, we'd better make sure it's conclusive. But many opinions we hold, theories we have and decisions we make are at least partially subjective – and that's fine, as long as we're honest about it, both with ourselves and with others. It's legitimate to have a position partly based on personal experience and gut feel as long as we don't claim that it's irrefutable. You might have a theory on the best way to raise children or develop a habit, and there's no problem if you acknowledge that it's only a conjecture, not a rule.

And even if the evidence *is* conclusive and we're firmly on the second rung from the top, evidence still has shortcomings. Chapter 6 highlighted how a regression allows us to include as many inputs as we'd like so we can control for dozens, even hundreds, of common causes – yet it can only ever feature a single output. The world's most rigorous regression can guide us on how best to pursue one unique goal, but most decisions in life involve several objectives.

If the only purpose of our company is to maximize profits, the sole reason we'd care about diversity is if it boosts the bottom line. Then, diversity advocates would put all their eggs into trying to prove that diversity improves performance, and see anyone who doubts this as a denier, perhaps insinuating sexism or racism. But if the goal of our business isn't just to make money but to contribute to a fairer and more equal society, then the link to profits is less important, because we can justify our diversity policy from a social angle. Evidence can never tell you what to do; it can only make us aware of the possible upsides and downsides to our actions. If the link to profits is negative, we then pursue the policy with our eyes open to the costs, rather than deceiving ourselves that it's a sure-fire road to riches.

Most choices we make are similarly multidimensional. You don't decide to play football rather than rugby because research demonstrates it has better fitness benefits. It might be that the football pitch is closer to home, your friends play for the same team – or you just love football. Nor do you enrol your kid in piano lessons over drama because they develop stronger cognitive skills, nor vote for Brexit or Remain purely based on the economic impact. Some of the most heated disagreements – whether between parents squabbling over how to bring up their children, executives debating a new strategy or citizens arguing which way to vote – arise because we have slightly different objectives. But because we never explicitly state our goals, we assume others have exactly the same ones as us, and so if we're right, they must be wrong. In fact, both positions could be justified, given our different ends.

And even if we have a single goal, evidence still can't dictate the only action we should take. Evidence only gives you an average result; it may not apply in every individual case, even considering the same context and range. What matters isn't whether the decision is right in general but whether it's right *for us*. Even if deliberate practice improves drumming better than jam sessions, and even if our only goal is rhythmic perfection rather than fun, we might just not like solo rehearsals and so they don't give us the same results as everyone else. Even if a low-carb diet were the best route to weight loss, if we have a craving for fruit, we won't be able to stick to it.

Understanding the limitations of evidence as well as its power helps us live more freely. A CEO can adopt a policy based on ethical principles, not just financial grounds. A dieter trying to lose weight can eat a banana and not feel guilty. The Olympic Games can showcase events ranging from javelin to gymnastics without taxpayers complaining that we’re not piling all their money into the single sport that best improves national health. I can bottle-feed our son if my wife is too exhausted to breastfeed and not worry about the IQ impact. We can have a discussion with someone of a different viewpoint without feeling threatened and be excited by the opportunity to learn rather than being afraid we’ll lose. We can enjoy and explore all the beautiful, complex and textured colours in issues often portrayed as black and white.

In a nutshell

- Creating societies that think smarter involves teaching critical thinking. This entails imparting:
    - Cognitive techniques, like ‘consider the opposite’: If a study found the opposite result, why might you be sceptical of it? (to address biased interpretation). What would you ask to disprove your theory? (to address biased search).
    - Statistical literacy, such as the existence of alternative explanations. This may be done through logic problems, such as the 2–4–6 task and the EK23 card game.
    - Curiosity through encouraging children to challenge and explore, and through films, TV programmes and public lectures on science, arts and the humanities.

    - Scaffolding: simple prompts such as ‘ensure you’ve given arguments for both sides’ and ‘check that all your points are relevant’.

- The *cultural cognition hypothesis* argues that people respond to a message based on the identity it portrays, not the evidence behind it. Public messaging should be disentangled from politics. For example, information on climate change could:
    - Be given by neutral parties, such as scientists, or by conservatives.
    - Refrain from ridiculing conservatives as climate-change deniers.
    - Emphasize that the solution requires conservative values, such as innovation.
- Books are rarely vetted by experts or even endorsers. Potential remedies are:
    - Centralized fact-checking websites for books, just as we have for quotes.
    - Frequent book endorsers publicly listing all books they have endorsed.
    - Publishers and retailers featuring a ‘health warning’ for disputed books.
- Journalists should refrain from citing a study unless they can link to it; researchers should not pitch a paper to the media unless they’ve made it freely available.

- On social media, fact-checking links reduce the sharing of fake news, but flagging posts as disputed means that unflagged ones are seen as more truthful.
- Citizens are not passive members of society; we shape society. Viewing dissenting arguments as an opportunity to learn, not something to argue against, contributes to a society where a diversity of viewpoints is embraced.
- Evidence is not an excuse for dogma. Two people may interpret data the same way but take different decisions due to contrasting objectives. Evidence only gives an average result; it doesn't apply in every setting. Understanding the limitations of evidence, as well as its power, helps us live more freely.

Appendix: A Checklist for Smarter Thinking

This Appendix provides a simple checklist that the reader can use to correctly evaluate statements, facts, data and evidence.

A. Preliminaries

1. *Do you want the conclusion to be true?*
2. *Is the conclusion extreme? Does it suggest that something is always good or always bad, or applies everywhere?*

   An example of a conclusion you might want to be true is ‘You can defeat cancer with diet’; an extreme conclusion is that ‘Carbs are always bad.’ If the answer to either question is yes, confirmation bias and/or black-and-white thinking may be at play. Not only might you accept the conclusion too readily, but the author may have deliberately skewed it to play to your biases. Then, it's particularly important to apply the following checks.

B. Statements

1. *Does the statement contain a superlative or imply universality?*

   An example of the former is 'Shareholder value is the dumbest idea in the world'; an example of the latter is 'Every company will be a fintech company.'
2. *If yes, can you come up with a clear counterexample?*

   If you can come up with worse ideas than shareholder value, or companies that might not become fintech firms, the statement is false and so you should put less weight on it. The author won't have literally meant that every single company will become a fintech company, but the extremism may have been used to mask the lack of actual evidence.
3. *Is the statement backed up by evidence, does the evidence exist, and is it publicly available?*

   Many statements claim 'There is clear evidence that . . .', but without citing anything. Sometimes an article might refer to evidence, even giving the authors' names, but the article is based on a press release when there's no study behind it. Or perhaps only an abridged version is available, which describes its results but not its methods, such as how it measured the input and output. If the full paper isn't public, much less weight should be put on it as there's no way to scrutinize its claims.

4. *If yes, does the evidence support the statement?*

   Sometimes the full paper is available, but it declares a result without actually showing it. The research claiming that high CEO pay discourages innovation gathered data on CEO contracts and assumed that they deter innovation, without testing the relationship. Alternatively, the evidence may contradict the claims, such as the paper claiming a link between diversity and performance when none of the ninety tests found one.

5. *Do the input and output correspond to the statement?*

   How are the input and output measured? This is particularly a concern if there's no clear way, such as for social distancing or long-term thinking. Be particularly suspicious if a measure is self-reported, such as perceived performance, or judged by the researcher, such as the *Built to Last* authors' own evaluation of whether a company followed the nine principles.

C. Facts

1. *Does the study test a hypothesis?*

Does it first form a hypothesis, such as ‘Starting with *why* leads to success,’ and then test it before drawing a conclusion? Or does it begin with its conclusion and then hand-pick examples consistent with it?

2. *Does the study consider a representative sample?*

   Does it include other people or companies with the same characteristic but different outcomes, such as companies that started with *why* but ended up unsuccessful?

3. *Does the study consider a control group?*

   Does it include people or companies without the characteristic but with the same outcome, such as companies that didn’t start with *why* but became successful?

4. *Does the study calculate the average output across the two groups?*

   Does it calculate the average output across not only the test group but also the control group? Is it upfront about the success of companies that didn’t start with *why*?

5. *Does the study check for statistical significance?*

   Does it test whether the difference between the two groups is large enough that it’s unlikely to be due to luck?

D. Data

1. *Are there other ways the researchers could have measured the input and the output?*

   If company performance is the output and it's measured using the profit margin, ask if there might be better metrics, such as shareholder returns. If diversity is the input and it's captured using the number of female directors, are there other reasonable diversity indicators they could have tried? If so, they may have data-mined their choice.

2. *Is the data chopped up?*

   Does the study divide the data into black-and-white buckets and ignore the full colour spectrum? For example, does it compare companies with at least three female directors to those with zero, ignoring the actual number? If so, the authors have thrown away data and the results may not hold using the full measures.

3. *Could the output have caused the input (reverse causation)?*

   If investment is correlated with future performance, it might be that when companies have good future prospects, they're more willing to invest – rather than investment causing the superior performance.

4. *Are there any common causes that could be driving both the input and output? Have the authors*

controlled for them in the same regression?

If a study claims that ‘People/companies that do X perform better,’ might people/companies who do X differ along many other dimensions? For example, mothers who breastfeed may have a more supportive home environment; CEOs who prioritize their employees’ emotions might be great leaders in other ways.

5. *Imagine the study found the opposite result. What alternative explanations would you come up with to try to explain it away? Then ask if these alternative explanations still apply, even though the findings are in the direction you like.*

   If a study finds that states with the death penalty have less crime than those without, a death-penalty advocate would lap it up. But if it had found a higher crime rate, he’d argue that other factors may be behind it, such as greater poverty. Now that he’s aware of alternative explanations, he should check if they apply even though the study finds the result he wants – perhaps the lower crime is due to less poverty.

E. Evidence

1. *What was the setting studied? Is this the same as the setting for which you’d like to draw conclusions? If not, are there any reasons why the relationship might*

be different in other settings?

Scientific management was shown to be successful in pig-iron handling, metal cutting and ball-bearing inspection. However, those settings involve one best way and a single measurable output. It may not apply to teaching, where there's no one best way and we care about multiple outputs, many of which aren't measurable.

2. *What was the population studied? Is this the same as the population of interest for which you'd like to draw conclusions? If not, are there any reasons why the relationship might be different in other populations?*

   Does the study consider only very successful people or companies? If so, even if it's controlled for common causes, it might be that the controls don't matter because they're already at very high levels and there are diminishing returns – fitness is immaterial for Beast Barracks recruits but a game-changer for the average person. Think about whether the controls might be relevant at more normal levels. Or, the input might only have an effect when the controls are high; for example, grit may only make a difference if you're exceptionally fit.

F. Shortcuts

If you don’t have the time or expertise to delve into a study and answer the questions in B to E above, a less thorough but faster approach is to ask the following instead:

Studies

1. Is the paper published in a top peer-reviewed journal?
2. What are the credentials of the authors? Do they have a Ph.D. and a track record of top peer-reviewed publications in the relevant field? Are they affiliated with a leading research institution? If the same study was written by the same authors, with the same credentials, but found the opposite results, would you still believe it?
3. What are the authors’ incentives to claim their result? Would they have published the paper if it had found the opposite result?
4. Do the authors exaggerate their credentials, the rigour of their methodology, or their conclusions?

Books and articles

1. Does the book or article back up its claims with evidence?
2. Can you find informed critiques on the web?
3. Is it balanced? Does it consider evidence or arguments

that contradict its core thesis?

4. Do the authors exaggerate their credentials, the rigour of their methodology, or their conclusions?

G. An example

Let’s put the above framework into practice by going through an example: the article ‘What if investors who held their shares longer got more voting power?’ by Roger L. Martin in *Harvard Business Review*. For brevity, we’ll consider only the opening paragraph, which is as follows:

> Joe Bower and Lynn Paine ‘had me at hello’ (to quote *Jerry Maguire*) with their new *HBR* article, ‘The Error at the Heart of Corporate Leadership.’ Laying out their data, they find that long-term oriented companies create more financial value and more jobs. In fact, if more American companies were focused on the long term, they estimate, investors would have an additional $1 trillion, workers would have an additional 5 million jobs, and the country would have more than an additional $1 trillion in GDP.

Starting with A, most readers would like this conclusion to be true – they think it’s good for companies to think long term. Indeed, Martin himself seems to suffer from confirmation bias, since he admits that the study ‘had me at hello’. It also makes extreme claims – if the world would only follow Bower and Paine’s advice in the article, $1 trillion of money and 5 million jobs would magically appear.

Moving to B, the paragraph does appear to be based on evidence – an article by Bower and Paine in *Harvard*

Business Review – but when you open it up, it’s on a completely different topic. However, there is an article by different authors in the same *HBR* issue that does make this claim – the McKinsey study ‘Finally, evidence that managing for the long term pays off ’, which we encountered earlier. One of their measures of long-termism is investment, but that’s problematic. High investment could be short-termist, like a football club spending millions in a desperate attempt to get into the Champions League.

Turning to C, the study does test a hypothesis: that long-termism leads to success. It also does consider the full population, of companies with both high and low investment. However, it doesn’t test for statistical significance, so the results might be due to luck.

On D, the researchers could have measured long-termism in other ways, such as whether the CEO is paid according to short-or long-term results. There are several common causes – a great CEO might invest more, as she has better ideas; and a great CEO could also directly improve company performance. Reverse causation is another problem, since good future prospects could have encouraged the company to increase investment.

Switching to E, the researchers find that companies that invest more perform better. Martin turns this description into a prediction, arguing that if low-investment companies increased their spending to the level of their open-handed peers, ‘investors would have an additional $1 trillion, workers would have an additional 5 million jobs, and the country would have more than an additional $1 trillion in GDP'. However, you can't apply the results for high-investment companies to low-investment ones. If businesses in declining industries, such as tobacco, invested more, they wouldn't perform better as they don't have good projects to invest in. The McKinsey study doesn't mention anything about investor returns, so Martin's statement that 'investors would have an additional $1 trillion' comes out of thin air.

Finally, on F, the McKinsey study isn't published in any peer-reviewed journal, nor is Martin's article – *Harvard Business Review* is a respected business magazine, but it isn't a scientific outlet with peer review. Martin himself is a prolific author, but with no Ph.D. or articles in top academic journals; he has expertise in management consulting but not data-driven research.

Acknowledgements

In keeping with its theme, this book benefited from a diversity of perspectives. I'd like to start with a big thank you to my agent Chris Wellbelove, commissioning editor Jamie Birkett and editor Celia Buzuk for their vision for this book, and in particular for constantly challenging me and recommending improvements when my biases led me to prefer the status quo. I thank Sarah Day for efficient copy-editing.

Writing a book on misinformation is dangerous, as you might fall foul of the mistakes that you describe. I thank researchers Jesús Romo, Aubrey Rugo and Sarah Qian for their invaluable help on specific sections, and Andrew Tickell, who combed through the entire text. London Business School and the Institute for Humane Studies generously provided funding for this research. Any remaining errors are, of course, my sole responsibility.

The manuscript was greatly enhanced by close readings, suggestions and constructive criticisms from Marc Canal, Chloe Fortier, Tom Gosling, Moqi Groen-Xu and Gaute Ulltveit-Moe. I continued to benefit from the mentorship of Will Hutton as I took the second step in my nascent career as an author. I thank everyone who filled in surveys and made suggestions on the book's title, in particular Connor Minney, whose advice proved pivotal, as well as those who responded to questions asking for examples of a particular bias or mistake.

The writing of this book was transformed by Lucy Emmerson and Clare Hayes Guymer. They turned stilted academic narrative into engaging stories, revamped my sentence and paragraph structures, clarified my lines of argument and probed my examples. This book would not be the same without them.

Notes

Introduction

[1] The full name was the Business, Energy and Industrial Strategy Committee.

[2] 'There is a lot of evidence that high inter-wage disparities within companies are detrimental to company performance.' Transcript of the 15 November 2016 oral evidence session as part of the Corporate Governance Inquiry (HC 702).

[3] Edmans, Alex (2011): 'Does the stock market fully value intangibles? Employee satisfaction and equity prices', *Journal of Financial Economics* 101, 621–40.

[4] Faleye, Olubunmi, Ebru Reis and Anand Venkateswaran (2010): 'The effect of executive–employee pay disparity on labor productivity'.

[5] Faleye, Olubunmi, Ebru Reis and Anand Venkateswaran (2013): 'The determinants and effects of CEO –employee pay ratios', *Journal of Banking and Finance* 37, 3258–72.

[6] The law required every UK listed company with over 250 employees to publish its pay ratio from 2020 (covering pay awarded in 2019).

[7] Ashworth-Hayes, Sam (2016): 'We don't send Brussels £350m a week', InFacts, 7 April 2016. Even the £120 million figure ignores any indirect benefits the UK received from EU membership, such as increased trade.

[8] Davenas, Elisabeth et al. (1988): 'Human basophil degranulation triggered by very dilute antiserum against IgE', *Nature* 333, 816–18.

[9] National Health and Medical Research Council (2015): ‘NHMRC information paper: evidence on the effectiveness of homeopathy for treating health conditions’, March 2015.

[10] Martin, Neil (2020): ‘Mars settlement likely by 2050 says UNSW expert – but not at levels predicted by Elon Musk’, *UNSW Newsroom*, 10 March 2021.

[11] Edwards, Erika and Vaughn Hillyard (2020): ‘Man dies after taking chloroquine in an attempt to prevent coronavirus’, NBC News, 23 March 2020.

[12] Santos, Laurie R. and Tamar Gendler (2014): ‘What scientific idea is ready for retirement? Knowing is half the battle’, Edge.org

[13] Walker, Mason and Katerina Eva Matsa (2021): ‘News consumption across social media in 2021’, Pew Research Center.

[14] Vosoughi, Soroush, Deb Roy and Sinan Aral (2018): ‘The spread of true and false news online’, *Science* 359, 1146–51.

1. Confirmation Bias

[*] Drilling mud is added as you dig to suspend the drilled-off rock so it floats to the surface, lubricate the drill bit, and – having twice the density of seawater – create pressure to prevent the well collapsing from the outside in.

[†] Functional magnetic resonance imaging.

[‡] As a result, the evidence we cite in this book will predominantly be from papers published in the top peer-reviewed journals, a point we’ll highlight in Chapter 9.

[§] For half the undergraduates, the first part of the experiment (selecting arguments) was on gun control and the second (evaluating arguments) on affirmative action, as we’ve described. For the other half, the first part was on affirmative action and the second on gun control.

[1] Belle was actually seventeen in 2009 but claimed to be twenty.

[2] Gibson, Belle (2014): *The Whole Pantry: Over 80 Original Gluten, Refined Sugar and Dairy-free Recipes to Nourish Your Body and Mind*. Penguin, Retrieved from https://books.google.co.uk/books?id=kdG9BAAAQBAJ

[3] Davey, Melissa (2016): ‘Belle Gibson video submitted to court sparks condemnation over cancer claims’, *Guardian*, 14 September 2016.

[4] Donelly, Beau and Nick Toscano (2017): *The Woman Who Fooled the World: Belle Gibson’s Cancer Con, and the Darkness at the Heart of the Wellness Industry*, Scribe Publications.

[5] Manavis, Sarah (2020): ‘How celebrities became the biggest peddlers of 5G coronavirus conspiracy theories’, *New Statesman*, 6 April 2020.

[6] Griffith, Erin (2021): ‘What red flags? Elizabeth Holmes trial exposes investors’ carelessness’, *New York Times*, 4 November 2021.

[7] Glover, Scott and Matt Lait (2005): ‘The evidence seemed overwhelming against Bruce Lisker but was justice served?’, *Los Angeles Times*, 22 May 2005.

[8] https://www.thesaurus.com/browse/be%20consistent%20with/. Accessed 23 September 2023.

[9] ‘The national registry of exonerations’. Available at https://www.law.umich.edu/special/exoneration/Pages/about.aspx

[10] Rossmo, D. Kim and Joycelyn M. Pollock (2019): ‘Confirmation bias and other systemic causes of wrongful convictions: a sentinel events perspective’, *Northeastern University Law Review* 11, 790–835.

[11] Nickerson, Raymond S. (1998): ‘Confirmation bias: a ubiquitous phenomenon in many guises’, *Review of General Psychology* 2, 175–220.

[12] Liu, Yao-Zhong et al. (2017): ‘Carcinogenic effects of oil dispersants: a KEGG pathway-based RNA-seq study of human airway epithelial cells’, *Gene* 602, 16–23.

[13] Denic-Roberts, Hristina et al. (2022): ‘Acute and longer-term cardiovascular conditions in the Deepwater Horizon oil spill coast guard cohort’, *Environment International* 158, 106937.

[14] Rusiecki, Jennifer A. et al. (2022): ‘Incidence of chronic respiratory conditions among oil spill responders: five years of follow-up in the Deepwater Horizon oil spill coast guard cohort study’, *Environmental Research* 203, 111824.

[15] ‘Report to the President’, National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, January 2011.

[16] Bartlit, Jr, Fred (2011): ‘Presidential oil spill commission releases report from chief counsel, Fred Bartlit’, Bartlit Beck LLP, February 2011.

[17] Gilbert, Daniel, Todd C. Frankel and Joseph Menn (2023): ‘Focused on profits, leaders made decisions that foreshadowed the bank’s surprise failure’, *Washington Post*, 2 April 2023.

[18] Kaplan, Jonas T., Sarah I. Gimbel and Sam Harris (2016): ‘Neural correlates of maintaining one’s political beliefs in the face of counterevidence’, *Scientific Reports* 6, 39589.

[19] Westen, Drew et al. (2006): ‘Neural bases of motivated reasoning: an fMRI study of emotional constraints on partisan political judgment in the 2004 US presidential election’, *Journal of Cognitive Neuroscience* 18, 1947–58.

[20] Lord, Charles G., Lee Ross and Mark R. Lepper (1979): ‘Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence’, *Journal of Personality and Social Psychology* 37, 2098–2109.

[21] Wason, Peter (1960): ‘On the failure to eliminate hypotheses in a conceptual task’, *Quarterly Journal of Experimental Psychology* 12, 129–40.

[22] Brock, Timothy C. and Joe L. Balloun (1967): ‘Behavioral receptivity to dissonant information’, *Journal of Personality and Social Psychology* 6, 413–28.

[23] Taber, Charles S. and Milton Lodge (2006). ‘Motivated skepticism in the evaluation of political beliefs’, *American Journal of Political Science* 501, 755–69.

2. Black-and-White Thinking

[*] Seniors are in the final year of secondary school (known as high school in the US).

[†] Phase 1 of the Atkins diet recommends no more than 20 grams of carbs per day. As a benchmark, a banana has 23 grams of carbs, although Phase 1 doesn’t allow any fruit at all.

[‡] Carnivores could be identified by sharp teeth and eyes at the front of the head.

[§] This point was made famous by the Australian comedy band The Axis of Awesome in their video ‘Four Chords’. In this video, they repeatedly play a progression of these four chords while singing lyrics from a medley of different songs, to highlight their ubiquity. The A, E and D chords are major; the F sharp chord is minor. Sometimes songs will branch out beyond those four but still stick within the key of A: they might throw in a B minor, C sharp minor, or – if they’re feeling particularly adventurous – a G sharp diminished. G is not a note in the key of A.

[‖] Conversely, for something that does good rather than creates harm, there’s no threshold that you need to hit (e.g. 10,000 hours of practice) for it to have value.

[#] These convert renewable forms of energy such as wind power into usable forms such as electricity.

[**] The researchers ran similar experiments on three other groups of students, giving them scientific reports, social judgements or consumer reviews instead of eyewitness testimonies, and found the same results.

[1] Rogak, Lisa (2005): *Dr. Robert Atkins: The True Story of the Man behind the War on Carbohydrates*, Chamberlain Bros.

[2] Gordon, Edgar S., Marchall Goldberg and Grace J. Chosy (1963): 'A new concept in the treatment of obesity', *Journal of the American Medical Association* 186, 50–60.

[3] Fisher, Roxanne (2013): 'What is the Atkins diet?', *BBC Good Food*, 24 September 2013.

[4] Trumbo, Paula et al. (2002): 'Dietary reference intakes for energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein and amino acid', *Journal of the American Dietetic Association* 102, 1621–30.

[5] Seidelmann, Sara B. et al. (2018): 'Dietary carbohydrate intake and mortality: a prospective cohort study and meta-analysis', *Lancet Public Health* 3, E419–E428.

[6] St. Jeor, Sachiko T. et al. (2001): 'Dietary protein and weight reduction: a statement for healthcare professionals from the Nutrition Committee of the Council on Nutrition, Physical Activity, and Metabolism of the American Heart Association', *Circulation* 104, 1869–74.

[7] DeLosh, Edward L., Jerome R. Busemeyer and Mark A. McDaniel (1997): 'Extrapolation: the sine qua non for abstraction in function learning', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23, 968–86.

[8] 'The world is going to miss the totemic 1.5°C climate target', *Economist*, 5 November 2022.

[9] Rozin, Paul et al. (1999): 'Individual differences in disgust sensitivity: comparisons and evaluations of paper-and-pencil versus behavioral measures', *Journal of Research in Personality* 33, 330–51.

[10] Rozin, Paul, Linda Millman and Carol Nemeroff (1986): 'Operation of the laws of sympathetic magic in disgust and other domains', *Journal of Personality and Social Psychology* 50, 703–12.

[11] Krueger, Joachim and Russell W. Clement (1994): 'Memory-based judgments about multiple categories: a revision and extension of Tajfel's accentuation theory', *Journal of Personality and Social Psychology* 67, 35–47.

[12] Cohen, Lauren, Umit G. Gurun and Quoc H. Nguyen (2021): 'The ESG-innovation disconnect: evidence from green patenting', NBER Working Paper 27990.

[13] Heeb, Florian et al. (2023): 'Do investors care about impact?', *Review of Financial Studies* 36, 1737–87.

[14] Fisher, Matthew and Frank Kiel (2018): 'The binary bias: a systematic distortion in the integration of information'. *Psychological Science* 29, 1846–58.

3. A Statement is Not Fact

[*] I wrote: ‘CEOs with high equity incentives outperform CEOs with low equity incentives by 4–10% per year, and the researchers do further tests to suggest that the results are causation rather than correlation.’

[†] Shareholder value is how much value a company creates for its shareholders. If the company is publicly traded, its market value (how much it costs on the stock market) is an estimate of shareholder value.

[‡] Reprinted from the *Journal of Financial Economics*, 1976, Michael C. Jensen and William H. Meckling, Theory of the Firm, Copyright 2023, with permission from Elsevier.

[§] The order of the names in the Wikipedia entry was corrected shortly after I wrote an article explaining what Jensen and Meckling actually wrote and highlighting Denning’s and Sinek’s misrepresentation.

[||] Milewski, Matthew D. et al., ‘Chronic lack of sleep is associated with increased sports injuries in adolescent athletes’, *Journal of Pediatric Orthopaedics* 34(2): 129–33, https://journals.lww.com/pedorthopaedics/fulltext/2014/03000/chronic_lack_of_sleep_is_associated_with_increased.1.aspx, with the first bar removed as in *Why We Sleep*.

[#] Milewski, Matthew D. et al., ‘Chronic lack of sleep is associated with increased sports injuries in adolescent athletes’, *Journal of Pediatric Orthopaedics* 34(2): 129–33, https://journals.lww.com/pedorthopaedics/fulltext/2014/03000/chronic_lack_of_sleep_is_associated_with_increased.1.aspx

[**] This was first pointed out by author Olli Haataja and popularized by researcher Alexey Guzey.

[††] EBITDA, or Earnings Before Interest, Tax, Depreciation and Amortization, is a measure of a company’s profits. The EBITDA margin is EBITDA divided by sales, which gauges the company’s profitability.

[‡‡] All companies in the UK have to pay the minimum wage set by law, known as the National Minimum Wage (or the National Living Wage for workers over twenty-three). The Living Wage (sometimes called the Real Living Wage to avoid confusion) is a suggested wage set by the Living Wage Foundation, a non-profit which recommends a higher wage based on the cost of living. Many companies voluntarily pay the Living Wage, but there is no legal obligation to do so.

[§§] For problems with the other studies cited, see ‘On ShareAction’s evidence in favour of the Sainsbury’s Living Wage resolution’ by Tom Gosling.

[||||] During the trial, the defence lawyer stated that Conservative peer Lord Astor denied having an affair with Rice-Davies, to which she famously replied, ‘Well he would, wouldn’t he?’ This is now commonly paraphrased as ‘He/she/they would say that, wouldn’t he/she/they?’

[##] The input and output are also known as the independent and dependent variables, respectively.

[***] Section 1233 is more than ten pages of dense text, which someone unfamiliar with medicine or the law might have difficulty in deciphering – unlike the sleep bar chart, which was the only graphic in a four-and-a-half-page article.

[1] Wong, Nathan Colin (2015): ‘The 10,000-hour rule’, *Canadian Urological Journal* 9, 299.

[2] Reingold, Jennifer (2008): ‘Secrets of their success’ (interview with Malcolm Gladwell), *Fortune*, 19 November 2008.

[3] Gladwell writes, ‘let’s test the [rule] with two examples’ and claims that the examples support the rule. Later on in the chapter, he writes, ‘there is an easy way to test this theory’, where the ‘theory’ refers to the hypothesis that you needed to be born in 1954 or 1955 to be a successful computer entrepreneur as you’d have accumulated 10,000 hours of practice by the start of the IT revolution.

[4] Edmans, Alex (2021): ‘What stakeholder capitalism can learn from Jensen and Meckling’, *ProMarket*, 9 May 2021.

[5] Milewski, Matthew D. et al. (2014): ‘Chronic lack of sleep is associated with increased sports injuries in adolescent athletes’, *Journal of Pediatric Orthopaedics* 34, 129–33.

[6] Two out of the ninety results were significant at the 10% level, but the standard threshold is 5%.

[7] Minerva Analytics (2021): ‘Boardroom diversity improves financial performance’, 23 July 2021.

[8] Dave, Dhaval M. et al. (2020): ‘Black Lives Matter protests, social distancing, and COVID-19’, IZA Institute of Labor Economics, Discussion Paper 13388.

[9] Thomas, Chloe et al. (2022): ‘The health, cost and equity impacts of restrictions on the advertisement of high fat, salt and sugar products across the Transport for London network: a Health economic modelling study’, *International Journal of Behavioral Nutrition and Physical Activity* 19, 93.

[10] Snowdon, Christopher (2022). 'Sadiq Khan is backing obesity claims based on shameless junk science', *Capx*, 2 August 2022.

[11] Rawson, Simon (2022): 'Living Wages for supermarket workers – decision time for investors', ShareAction, 4 July 2022.

[12] Heery, Edmund, David Nash and Deborah Hann (2017): 'The Living Wage employer experience', Cardiff Business School.

[13] McKinsey & Company (2020): 'COVID-19: briefing note', 25 March 2020.

[14] Bolger, Thomas et al. (2019): 'The invisible drag on UK R&D: how corporate incentives within the FTSE 350 inhibit innovation', Nesta, 7 August 2019.

[15] Urso, Federica and Simon Jessop (2022): 'Boardrooms with more women deliver more on climate, says Arabesque', Reuters, 22 March 2022.

[16] Pew Research Center (2009): 'Health care reform closely followed, much discussed', 20 August 2009.

[17] Fancy, Tariq (2021): 'The secret diary of a "sustainable investor"', Medium, 20 August 2021.

[18] Edmans, Alex (2021): 'Is sustainable investing really a dangerous placebo?', Medium, 30 September 2021.

[19] Power, William (2021): 'Does sustainable investing really help the environment?', *Wall Street Journal*, 7 November 2021.

[20] Eccles, Robert G. (2022): 'The topology of hate for ESG', *Forbes*, 3 June 2022.

[21] 'John Mearsheimer on why the West is principally responsible for the Ukrainian crisis', *Economist*, 19 March 2022.

4. A Fact is Not Data

[*] It could be that each theory offers a partial explanation, so they're not inconsistent. However, the inconsistency arises as both authors imply they've explained everything behind Apple's success and don't acknowledge the possibility of rival theories.

[†] They defined 'frequent' traders as those in the top 20% of trading frequency.

[‡] The calculation is 1.03311.

[§] Formally, a statistical test is of a *null hypothesis* that there is no relationship, i.e. 'Frequent trading has no effect on returns.' 'Frequent trading affects returns' is known as the *alternative hypothesis*, and the test pits the null hypothesis against the hypothesis. When we use the word 'hypothesis' in this book, it refers to alternative hypotheses, since most of the time we conjecture that a relationship exists. In addition, this book uses ‘alternative hypothesis’ to refer to something else – even if there is a relationship, there are multiple possible explanations for it, and an ‘alternative hypothesis’ is a rival theory for your chosen explanation. This is the common use of the phrase in practice.

[||] The more technical term is ‘treated sample’, because it’s been ‘treated’ by the input we’re interested in – in this case, frequent trading.

[#] Brad and Terry’s paper doesn’t actually calculate the statistical significance for the difference between 17.9% and 11.4%. This is because they don’t draw or claim any inferences from this difference, as it doesn’t address alternative explanations – the topic of Chapter 6. They conduct other analyses to address alternative explanations, and then conduct formal significance tests on these results.

[**] The probability of red or black twenty-six times straight is $(18/37)^{26-1}$.

[††] Note that even a representative sample would not be able to *prove* that Jobs’s adoption led to his success. As we’ll stress in Chapter 8, evidence is not proof – it only finds an average relationship, which may not hold in every setting. Even if adopted CEOs perform better than non-adopted ones on average, this need not mean that adoption improved performance in Jobs’s case.

[‡‡] You’d then need to do the same for those who chose chemotherapy and radiotherapy.

[§§] This quote is from Isaacson’s book, but the book then quickly forgets Jobs’s own claim that he never felt abandoned. The following section describes how Steve’s adoptive parents cared for him and concludes: ‘So he grew up not only with a sense of having once been abandoned, but also with a sense that he was special.’

[||||] We would also need a control group of non-biochemistry graduates, both those who ended up wealthy and those who didn’t.

[##] A separate problem is that there’s no objective way to measure whether the successful companies followed the nine principles and the unsuccessful ones didn’t. This is similar to the concern of data being self-reported in Chapter 3, but here the data is reported by the researchers, not the companies. For example, the principle of ‘Good enough never is’ states that the companies should always strive to improve, but you can’t measure efforts to improve, only actual improvement. Two marathon runners could put in the same effort even if they have different times. The reader has to rely on Collins and Porras’s assessment, and they can always claim that leaders followed the principles and laggards didn’t because this assessment is subjective.

[1] Issacson, Walter (2011): *Steve Jobs*, Simon & Schuster.

[2] While citizens have to disclose gains and losses in their tax returns, many may be below the filing threshold (e.g. in the UK, you don’t have to report if your net gains are below an allowance, currently £6,000). In addition, tax returns are confidential and researchers don’t have access to them.

[3] Barber, Brad and Terrance Odean (2000): ‘Trading is hazardous to your wealth: the common stock investment performance of individual investors’, *Journal of Finance* 55, 773–806.

[4] Curry, Colleen (2013): ‘Jeff Bezos and Steve Jobs: both estranged from dads and wild tech successes’, ABC News, 12 October 2013.

[5] Isaacson, Walter (2012): ‘The real leadership lessons of Steve Jobs’, *Harvard Business Review*, April 2012.

[6] Staw, Barry M. (1975): ‘Attribution of the “causes” of performance: a general alternative interpretation of cross-sectional research on organizations’, *Organizational Behavior and Human Performance* 13, 414–32.

5. Data is Not Evidence: Data Mining

[*] Only the names of the Best Companies are published, not their individual scores.

[†] These were either measures exclusively on gender diversity in particular, or on diversity in general (since gender diversity is a key component of overall diversity). We excluded measures focused on other aspects of diversity, such as ethnic diversity.

[‡] The probability of six heads *or* six tails is 3.125%, which is below 5%.

[§] Recall that for a link to be statistically significant, the likelihood it arose from pure chance must be 5% or less. Thus, there’s a 5% likelihood that one test yields a significant result due to luck. If you run a hundred independent tests, on average 5% × 100 = 5 will be significant.

[||] Out of 20 tests, on average one will uncover a significant link in either the positive or negative direction. If you want a positive and significant result, on average you’ll need to run 40 tests.

[#] Note this is a quite different problem to what we saw in Chapter 3, where studies used measures that had little to do with what they claimed. Here, all twenty-four were valid ways to gauge diversity, but we had the freedom to pick and choose the ones that worked.

[**] Shareholder returns are the change in the stock price, plus dividends.

[††] Specifically, they used three measures of the profit margin: return on sales, return on invested capital and return on equity. The problems apply to all three measures of the profit margin.

[‡‡] More precisely, statistical significance means that there's a 5% probability that a result is due to luck if it's the only test that you ran. However, if you ran multiple tests and reported the highest result, then the probability that it's due to luck is much higher than 5% – meaning that it's much less likely there's an actual relationship.

[§§] National (American) Football League.

[||||] There might be justifiable reasons for why you should focus on data post-2007. Perhaps the world changed after 2007 and so what happened before isn't so relevant today. If so, Thomson Reuters could have presented all the data, then cut it into pre-2007 and post-2007, and explained why they thought 2007 was the relevant breakpoint. The reader could then reach her own conclusion about whether to focus only on the more recent period or to consider the full history. But hiding key data to prevent the reader thinking for herself and drawing her own implications is misleading.

[##] We only studied international sports, since it's clear how they'll affect the stock market. With regional competitions, if Manchester City win and Liverpool lose, then some Brits are elated and others devastated, so it's unclear what should happen to the UK stock market.

[***] More precisely, the best-fit line minimizes the sum of the squared distances between each point and the line. You don't need to draw a graph to run a regression; you just put the numbers into a formula and it gives you the slope. 'Slope' and 'gradient' are typically used to describe the steepness of the best-fit line in a graph; when calculated from a set of numbers without a graph, it's often referred to as a *regression coefficient*.

[†††] The more technical terms are *discretization* or *binarization.* They are examples of a more general way to engage in data mining, which is to choose your *specification* – how to combine the data that you have into an overall result. For brevity, we focus only on grouping. However, there are many other specification choices that researchers can make to data-mine. Moreover, our example is one where the input is grouped, but researchers can also data-mine by grouping the output, for instance by studying how diversity affects whether profits are in the top or bottom half, or the top or bottom third, rather than the actual level of profits.

[‡‡‡] The researcher does have freedom to make other specification choices beyond the scope of this book. Given a set of data, there is only one way to run a standard regression.

[§§§] Note that grouping can sometimes be justified. Perhaps female directors only boost company performance once they reach a certain number (such as three) since you need critical mass – a lone female can't achieve much. However, if so, the researchers should provide a clear justification for grouping, and check that the results are robust to other groupings, to reassure the reader that they haven't fished for the one categorization that works.

[1] The £800 figure is in today's prices.

[2] Edmans, Alex (2011): 'Does the stock market fully value intangibles? Employee satisfaction and equity prices', *Journal of Financial Economics* 101, 621–40.

[3] Hill, Russell A. and Robert A. Barton (2005): 'Red enhances human performance in contests', *Nature* 435, 293.

[4] Elliot, Andrew J. et al. (2007): 'Color and psychological functioning: the effect of red on performance attainment', *Journal of Experimental Psychology: General* 136, 154–68.

[5] Kamstra, Mark J., Lisa A. Kramer and Maurice D. Levi (2000): 'Losing sleep at the market: the daylight saving anomaly', *American Economic Review* 90, 1005–11.

[6] Kamstra, Mark J., Lisa A. Kramer and Maurice D. Levi (2003): 'Winter blues: a SAD stock market cycle', *American Economic Review* 93, 324–43.

[7] Hirshleifer, David and Tyler Shumway (2003): 'Good day sunshine: stock returns and the weather', *Journal of Finance* 58, 1009–32.

[8] Yuan, Kathy, Lu Zheng and Qiaoqiao Zhu (2006): 'Are investors moonstruck? Lunar phases and stock returns', *Journal of Empirical Finance* 13, 1–23.

[9] Carroll, Douglas et al. (2002): 'Admissions for myocardial infarction and World Cup football: database survey', *British Medical Journal* 325, 1439–42.

[10] Trovato, Frank (1998): 'The Stanley Cup of hockey and suicide in Quebec, 1951–1992', *Social Forces* 77, 105–26.

[11] White, Garland F. (1989): 'Media and violence: the case of professional football championship games', *Aggressive Behavior* 15, 423–33.

[12] Edmans, Alex, Diego García and Øyvind Norli (2007): 'Sports sentiment and stock returns', *Journal of Finance* 62, 1967–98.

[13] Chanavat, André and Katharine Ramsden (2013): 'Mining the metrics of board diversity', Thomson Reuters.

[14] ‘The effect of the 2014 World Cup on stock markets – Alex Edmans and CNN’s Richard Quest’. Available at https://bit.ly/soccercnn

6. Data is Not Evidence: Causation

[*] The ‘p-value’ corresponds to the significance level, which needs to be 0.05 or lower for a result to be deemed significant.

[†] More technical terms for ‘common causes’ are ‘omitted factors’, ‘omitted variables’ or ‘confounding variables’.

[‡] Ideally, we’d have read all the required papers and books before the birth. Instead, we’d gone by conventional wisdom, the NCT course and my desktop search, and planned to exclusively breastfeed Caspar. But, as Mike Tyson said, ‘Everyone has a plan until you get punched in the mouth.’

[§] For the short-term benefits to the baby, there are fewer allergic rashes, gastrointestinal disorders and possibly ear infections, and a lower chance of necrotizing enterocolitis (a serious intestinal infection). Mums have less risk of breast cancer.

[‖] This paper actually had no control group – it just showed that subjects who followed the diet lost weight. Even if it had a control group of nondieters and showed that their weights didn’t change, it still couldn’t conclude that the diet caused the weight loss for the reasons given in the text.

[#] One component of trust was stable financial performance. It’s almost tautological that companies with stable financial performance will perform better; this has nothing to do with trust.

[**] The post also suffered from problems that we haven’t got to yet, such as it being almost impossible to show proof (Chapter 8).

[††] For example, *Harvard Business Review* has articles entitled ‘Companies that practice “conscious capitalism” perform 10x better’, ‘Companies that invest in sustainability do better financially’ and ‘CEOs with diverse networks create higher firm value’.

[‡‡] There were many possible measures of employee satisfaction, so how did I convince readers I hadn’t mined for one that worked? The list was the only measure I tried for two reasons. It’s been around since 1984, so I had 28 years of data, not just a single data point (my paper was published in 2012 so my data stopped in 2011). It’s also extremely thorough, surveying 250 employees per company on issues spanning credibility, fairness, respect, pride and camaraderie.

[§§] The calculation is $1.023^{28–1}$.

[||||] Technically, this is known as *multivariate regression* since it uses multiple inputs, but it’s often just called a regression.

[##] The full term is ‘control variables’.

[***] After adding all those extra factors, the outperformance of the Best Companies was now 4.1% per year. Note that this is not directly comparable to the 2.3% per year previously mentioned, because a different regression technique is used when there are many controls.

[†††] While excluding the control won’t affect the coefficient (slope), it will lower the statistical significance. Thus, if you already have a significant result, it can’t be attacked by your failure to control for a factor that’s unrelated to the input, because if you did control for it, the significance would be even higher.

[1] Castro, Rita Amiel et al. (2021): ‘Breastfeeding, prenatal depression and children’s IQ and behaviour: a test of a moderation model’, *BMC Pregnancy and Childbirth* 21, 62.

[2] Al Thuneyyan, Danyah Abdullah et al. (2022): ‘The effect of breastfeeding on intelligence quotient and social intelligence among seven-to nine-year-old girls: a pilot study’, *Frontiers in Nutrition* 9, 726042.

[3] Bayer, Ilker S. (2018): ‘Advances in tribology of lubricin and lubricin-like synthetic polymer nanostructures’, *Lubricants* 6, 3; Jay, Gregory D. and Kimberly A. Waller (2014): ‘The biology of lubricin: near frictionless joint motion’, *Matrix Biology* 39, 17–24. The first article gives synovial fluid’s friction coefficient as 0.001; the second article gives Teflon’s as 0.04.

[4] Der, Geoff G., David Batty and Ian J. Deary (2006): ‘Effect of breast feeding on intelligence in children: prospective study, sibling pairs analysis, and meta-analysis’, *British Medical Journal* 333, 945.

[5] EY (2022): ‘Press release: prioritizing emotions is the key to success for business transformation’, 28 June 2022.

[6] Saïd Business School (2022): ‘Prioritising emotions is the key to success for business transformation, groundbreaking research finds’, 28 June 2022.

[7] Travaglio, Marco et al. (2020): ‘Link between air pollution and Covid-19 in England’, *Environmental Pollution* 268, 115859.

[8] Carrington, Damian (2020): ‘“Compelling” evidence air pollution worsens coronavirus – study’, *Guardian*, 13 July 2020.

[9] Barton, Dominic, James Manyika and Sarah Keohane Williamson (2017): ‘Finally, evidence that managing for the long term pays off’, *Harvard Business Review*, 7 February 2017.

[10] Edmans, Alex (2012): ‘The link between job satisfaction and firm value, with implications for corporate social responsibility’, *Academy of Management Perspectives* 26, 1–19.

[11] Rambotti, Simone (2015): ‘Recalibrating the spirit level: an analysis of the interaction of income inequality and poverty and its effect on health’, *Social Science & Medicine* 139, 123–31.

[12] Wei, Yuan et al. (2021): ‘Smoking cessation in late life is associated with increased risk of all-cause mortality amongst oldest old people: a community-based prospective cohort study’, *Age and Ageing* 50, 1298–1305.

7. When Data is Evidence

[*] This figure is taken from the Caroline Hoxby paper that we will discuss shortly. Travel times may have changed since she wrote the paper.

[†] School districts compete with each other for students, because budgets are determined at district rather than school level. One reason is that some programmes (e.g. for disabled children) are run at district rather than school level.

[‡] More specifically, nothing is causing the input beyond luck (e.g. a card draw). If the result is statistically significant, this means that it’s very unlikely that the output was driven by luck, and so luck isn’t a common cause.

[§] Note that differences in education and work experience may be themselves due to discrimination. However, this explanation would suggest different remedies to the case in which African Americans and Caucasians with the *same* credentials are treated differently.

[||] These jobs were sales, administrative support, clerical services, and customer services.

[#] The researchers then added extra features such as language or computer skills. This accentuated the differences in quality, avoiding the problem that some CVs were in the middle and could have been classified either way; it also made the résumés distinct and so didn’t compromise their actual owners.

[**] The more technical term is *instrumental variable*.

[††] The exogenous part is also known as the ‘explained’ or ‘instrumented’ component, and the endogenous part as the ‘unexplained’ component. You calculate the ‘explained’ part by running a regression of the output on the instrument; this shows you how much the instrument can explain the output. If there are observable common causes that affect the output, then these are included in the ‘explained’ part alongside the instrument. When estimating the

‘explained’ component of school choice, Hoxby included factors such as the population, average income and racial composition of the metropolitan area together with the number of rivers. The common causes in the diagram are unobservable ones.

‡‡ For example, consider a company that’s the only employer in its industry in a particular region (known as a ‘local monopsony’). Since it’s the only employer, it can set its own wages, rather than having to pay what competitors are paying. Assume that, to hire ten workers, it can offer a wage of $15/ hour, but to hire eleven, it needs to offer $16/hour (to pull that extra worker in from another industry). If employees produce output of $20 an hour, it may seem worth it to hire eleven workers, as the gain from hiring that eleventh worker ($20) exceeds the cost ($16). However, the cost isn’t just the $16 you pay to the eleventh worker but the extra $1/hour that you’d need to pay the other ten workers, and so the firm will only hire ten workers. However, a minimum wage of $16 will lead to the company voluntarily choosing to hire eleven workers – it will have to pay the first ten workers $16 anyway, so it doesn’t need to pay them more if it hires the eleventh worker.

§§ The data has been rounded to the nearest whole number for ease of exposition.

|||| How do natural experiments and instruments compare? They’re similar in that both are shocks that approximate the random assignment in an RCT. The difference is that, with a natural experiment, it’s the input we’re interested in that varies randomly. In the above example, we’re interested in the minimum wage, and the law change directly shocked the minimum wage. An instrument also varies randomly, but we’re not interested in the instrument itself but the input it shocks. In the school choice study, we care about how child performance is affected by school choice, not the number of rivers; in the family succession research, we’re interested in how profitability depends on the CEO, not the predecessor’s first-born child.

## Under our emotional explanation, shouldn’t pre-game odds also matter – an unexpected loss affects investor mood more than an expected loss? Actually, this needn’t be the case – an England loss to Germany is more expected than a loss to Iceland, but the former still hurts because Germany is a historic rival.

*** More precisely, employee satisfaction causes the stock price to go up because it improves profits and the market didn’t fully anticipate this improvement, as demonstrated by analysts under-forecasting earnings. Then, when earnings come in higher than expected, the stock price jumps.

[1] Bown, Stephen R. (2003): *Scurvy: How a Surgeon, a Mariner, and a Gentleman Solved the Greatest Medical Mystery of the Age of Sail*, Summersdale.

[2] Bureau of Labor Statistics (2023): ‘Usual weekly earnings of wage and salary workers second quarter 2023’, 18 July 2023. In Q3 2022, the median weekly earnings of African Americans in the US was $881 compared to $1,101 for Caucasians.

[3] Bertrand, Marianne and Sendhil Mullainathan (2004): ‘Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination’, *American Economic Review* 94, 991–1013.

[4] Hoxby, Caroline M. (2000): ‘Does competition among public schools benefit students and taxpayers?’, *American Economic Review* 90, 1209–38.

[5] Bennedsen, Morten et al. (2007): ‘Inside the family firm: the role of families in succession decisions and performance’, *Quarterly Journal of Economics* 122, 647–91.

8. Evidence is Not Proof

[*] The standards for evidence were lower in Taylor’s day, as modern techniques such as instruments hadn’t yet been developed. But even if Taylor conducted his experiments with the utmost precision and had the perfect control group, they would only be evidence, not proof.

[†] More precisely, grit combines perseverance of effort (working hard in the face of setbacks) with consistency of interests (not changing one’s goals). Duckworth refers to the latter as ‘passion’ for short.

[‡] The SAT, originally called the Scholastic Aptitude Test, is a standardized test used for admission to US universities.

[1] Taylor, Frederick Winslow (1907): *On the Art of Cutting Metals*, American Society of Mechanical Engineers.

[2] Taylor, Frederick W. (1911): *The Principles of Scientific Management*, Harper & Brothers.

[3] Au, Wayne (2011): ‘Teaching under the new Taylorism: highstakes testing and the standardization of the 21st century curriculum’, *Journal of Curriculum Studies* 43, 25–45.

[4] Bobbitt, John Franklin (1912): ‘The elimination of waste in education’, *Elementary School Teacher* 12, 259–71.

[5] Ireh, Maduakolam (2016): ‘Scientific management still endures in education’. Available at https://files.eric.ed.gov/fulltext/ED566616.pdf

[6] Paige, Rod (2003): Letter to the editor, *New Yorker*, 6 October 2003.

[7] Center on Education Policy (2006): ‘From the capital to the classroom: year 4 of the No Child Left Behind Act’.

[8] Houghton Mifflin Reading: *A Legacy of Literacy*, California Teacher’s Edition, grade 1.

[9] Iasevoli, Brenda (2017): ‘Teachers go public with their resignation letters’, *Education Week*, 14 April 2017.

[10] Ferriss, Timothy (2010): *The 4-Hour Body: An Uncommon Guide to Rapid Fat-loss, Incredible Sex, and Becoming Superhuman*, Crown Publishing Group.

[11] Duckworth, Angela (2016): *Grit: The Power of Passion and Perseverance*, Simon and Schuster, New York.

[12] Duckworth, A. L. et al. (2007): ‘Grit: perseverance and passion for long-term goals’, *Journal of Personality and Social Psychology*, 92(6), 1087–101.

[13] Duckworth, A. L. et al. (2011): ‘Deliberate practice spells success: why grittier competitors triumph at the National Spelling Bee’, *Social Psychological and Personality Science*, 2(2), 174–81.

[14] Duckworth, A. L. et al. (2007): ‘Grit: perseverance and passion for long-term goals’, *Journal of Personality and Social Psychology*, 92(6), 1087–101.

[15] Scelfo, Julie (2016): ‘Angela Duckworth on passion, grit, and success’, *New York Times*, 8 April 2016.

[16] Credé, Marcus (2018): ‘What shall we do about grit? A critical review of what we know and what we don’t know’, *Educational Researcher* 47, 606–11; Credé, Marcus, Michael C. Tynan and Peter D. Harris (2017): ‘Much ado about grit: a meta-analytic synthesis of the grit literature’, *Journal of Personality and Social Psychology*, 113(3), 492–511.

[17] Sawka, Michael N. et al. (2007): ‘American College of Sports Medicine Position Stand: exercise and fluid replacement’, *Medicine and Science in Sports and Exercise*, 39, 377–90.

[18] Yeh, Robert W. et al. (2018): ‘Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial’, *British Medical Journal*, 363.

9. Thinking Smarter as Individuals

[*] If the person is being offensive or posting misinformation, this is a justification to unfollow them.

[†] Such learning is known as ‘Bayesian updating’, and what is updated is your initial belief.

[‡] This is known as a ‘replication study’. Replication studies typically try to make improvements on the original experimental design, for example using a larger number of participants and ensuring that the control group has a placebo.

[§] The retraction was because Wakefield flouted ethics guidelines while performing his research; a year later, his results were shown to be fraudulent.

[||] The *American Economic Review*, which published Caroline Hoxby’s paper on school choice and child performance, later published a Comment by Jesse Rothschild arguing that the results become insignificant under different methodologies, as well as a Reply by Hoxby arguing that the alternative approaches are invalid or don’t affect the results.

[#] Books published by trade presses, which typically have the most widespread audiences, are never peer-reviewed. Books published by university presses are peer-reviewed, but the review is much more superficial and is mainly on the novelty, structure and general thesis of the book rather than the evidence used. The editor who makes the publication decision is not an academic, unlike for an academic journal.

[**] In physical sciences and medicine, evidence can be unambiguous, so balance needn’t be a necessary criterion.

[††] This occurs in a play-within-the-play where the actress portraying Queen Gertrude, Hamlet’s mother, claims that she won’t remarry if her husband, the King of Denmark, dies. Her over-the-top language makes her appear insincere, leading the real Queen Gertrude to quip, ‘The lady doth protest too much, methinks.’

[‡‡] Professor is a position at a university that involves both research and teaching, while a doctorate is a qualification. People may have doctorates but may choose not to work at a university.

[§§] An adjunct professor is a part-time lecturer with no research requirements. ‘Practice professor’, ‘clinical professor’ and ‘visiting professor’ are similar titles. An honorary doctorate is awarded without any study or research; instead it’s given for other significant achievements. For example, Muhammad Ali, Aretha Franklin and Kanye West have honorary doctorates; P. Diddy has one from Harvard; and Meryl Streep has five.

[||||] However, a claim such as ‘*New York Times*/*Sunday Times* bestselling author’ is specific and verifiable. All books referred to as ‘bestselling’ in this book are *New York Times* bestsellers.

[##] The claim is valid if a specific metric is given, such as citations.

[1] Covey, Stephen R. (1989): *The 7 Habits of Highly Effective People*, Free Press.

[2] Securities and Exchange Commission 17 CFR Parts 229 and 249 (Release Nos. 33-9877; 34-75610; File No. S7-07-13).

[3] Ioannidis, John P. A. (2015): ‘Stealth research: is biomedical innovation happening outside the peer-reviewed literature?’, *Journal of the American Medical Association*, 313, 663–4.

[4] Carreyrou, John (2018): *Bad Blood: Secrets and Lies in a Silicon Valley Startup*, Penguin Random House.

[5] Emerson, Gwendolyn B. et al. (2010): ‘Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial’, *Archives of Internal Medicine* 170, 1934–9.

[6] Carney, Dana R., Amy J. C. Cuddy and Andy J. Yap (2010): ‘Power posing: brief nonverbal displays affect neuroendocrine levels and risk tolerance’, *Psychological Science* 21, 1363–8.

[7] Ranehill, Eva et al. (2015): ‘Assessing the robustness of power posing: no effect on hormones and risk tolerance in a large sample of men and women’, *Psychological Science* 26, 652–6.

[8] Wakefield, A. et al. (1998): ‘Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children’, *Lancet* 351 (9103): 637–41.

[9] National Health and Medical Research Council (2015): ‘NHMRC information paper: evidence on the effectiveness of homeopathy for treating health conditions’, March 2015.

[10] Grant Thornton (2019): ‘Corporate governance and company performance: a proven link between effective corporate governance and value creation’.

[11] Fabo, Brian et al. (2021): ‘Fifty shades of QE: comparing findings of central bankers and academics’, *Journal of Monetary Economics* 120, 1–20.

[12] Allen, David (2001): *Getting Things Done*, Penguin.

[13] Keegan, Paul (2007): ‘How David Allen mastered getting things done’, *Business 2.0 Magazine*, 21 June 2007.

[14] Hatmaker, Taylor (2010): ‘Twitter plans to bring prompts to “read before you retweet” to all users’, *TechCrunch*, 24 September 2020.

[15] Pennycook, Gordon et al. (2021): ‘Shifting attention to accuracy can reduce misinformation online’, *Nature* 592, 590–95.

10. Creating Organizations that Think Smarter

[*] Demographic diversity could be pursued for reasons other than overcoming groupthink, such as social justice.

[†] Sample tasks include: brainstorming new ideas (on the possible uses of a brick), making moral judgements (deciding disciplinary action in a fictitious case where a college athlete bribes an instructor to change his grade), allocating limited resources (planning a group shopping trip where each member wants different groceries) and coordination (typing a text into a shared document, where members work independently and simultaneously and so have to avoid duplicating colleagues or missing words).

[‡] The same results held when studying sorority sisters.

[§] The blockade was technically referred to as a 'quarantine', since a blockade is officially an act of war.

[||] Amazon founder Jeff Bezos uses the term 'study hall'.

[#] They were first introduced by the Catholic Church, in response to concerns that people canonized as saints weren't actually that deserving; the devil's advocate was tasked with finding flaws in the candidate.

[**] Red teams were initially used to simulate the enemy in military wargaming, to hypothetically attack the home 'blue team'. They're also used in other settings with a defined adversary, such as appointing hackers to try to break into a cybersecurity system, undercover agents tasked with getting deadly weapons through airport security, or lawyers playing the role of opposing counsel to tear apart your case. However, even if there is no known enemy, a team can work together to identify the risks of a strategy.

[††] Examples include building defences against sea-level rise and developing crops that grow in warmer climates.

[1] Janis, Irving L. (1971): 'Groupthink', *Psychology Today* 5, 84–90.

[2] Malmendier, Ulrike (2021): 'Experience effects in finance: foundations, applications, and future directions', *Review of Finance* 25, 1339–63.

[3] PwC and AIESEC (2016): 'Tomorrow's leaders today'.

[4] Dallek, Robert (2013): 'JFK vs. the military', *Atlantic*, 10 September 2013.

[5] Aggarwal, Ishani et al. (2019): 'The impact of cognitive style diversity on implicit learning in teams', *Frontiers in Psychology* 10, 112.

[6] Fos, Vyacheslav, Elisabeth Kempf and Margarita Tsoutsoura (2023): 'The political polarization of corporate America', National Bureau of Economic Research 30183.

[7] Loyd, Denise Lewin et al. (2013): ‘Social category diversity promotes premeeting elaboration: the role of relationship focus’, *Organization Science* 24, 757–72.

[8] Phillips, Katherine W., Katie A. Liljenquist and Margaret A. Neale (2009): ‘Is the pain worth the gain? The advantages and liabilities of agreeing with socially distinct newcomers’, *Personality and Social Psychology Bulletin* 35, 336–50.

[9] Edmans, Alex, Caroline Flammer and Simon Glossner (2023): ‘Diversity, equity, and inclusion’. NBDR Working Paper 31215.

[10] Allison, Graham T. and Philip D. Zelikow (1999): *Essence of Decision: Explaining the Cuban Missile Crisis* (2nd edn), Longman, pp. 111–16.

[11] Bikhchandani, Sushil, David Hirshleifer and Ivo Welch (1992): ‘A theory of fads, fashion, custom, and cultural change as informational cascades’, *Journal of Political Economy* 100, 5, 992–1026.

[12] Soeters, Joseph L. and Peter C. Boer (2000): ‘Culture and flight safety in military aviation’, *International Journal of Aviation Psychology* 10, 111–33.

[13] Rozenblit, Leonid and Frank Keil (2002): ‘The misunderstood limits of folk science: an illusion of explanatory depth’, *Cognitive Science* 26, 521–62.

[14] Fernbach, Philip M. et al. (2013): ‘Political extremism is supported by an illusion of understanding’, *Psychological Science* 24, 939–46.

11. Creating Societies that Think Smarter

[*] The problem is as follows: You have a fox, a chicken and a sack of grain. You must cross a river with only one of them at a time. If you leave the fox with the chicken he will eat her; if you leave the chicken with the grain she will eat it. How can you get all three across safely?

[†] The researchers compared people with ‘hierarchical’ and ‘individualistic’ world views against those with ‘egalitarian’ or ‘communitarian’ beliefs. For brevity, I refer to the former as right wing and the latter as left wing.

[‡] For example, one expert was said to have written the book *Three Social Evils: Sexism, Racism, and Homophobia*, and another *The Immigrant Invasion: Threatening the American Way of Life*. The former would likely be viewed as left wing and the latter right wing.

[§] TED also changed the title from ‘Your body language shapes who you are’ to ‘Your body language may shape who you are’.

[||] The first claim was that 87% of French laws come from European directives; the second stated that the EU wants to attract 50 million immigrants to Europe by 2050.

[#] You might wonder how best practices can make a difference if regulation can't. With regulation, the burden of proof to prosecute someone for misinformation is very high. For best practices, it's much lower; someone who consistently violates best practices can be censured even if they cannot be prosecuted.

[**] One book that has commented on the limited signal provided by endorsements is Nassim Taleb's *Fooled by Randomness*. Taleb highlights selection bias – publishers reach out to many potential endorsers and only include the most positive ones. Here, we highlight a separate issue – even if publishers included all endorsements, they'd be overly positive as there's no deterrent to inflation.

[1] Lord, Charles G., Mark R. Lepper and Elizabeth Preston (1984): 'Considering the opposite: a corrective strategy for social judgment', *Journal of Personality and Social Psychology* 47, 1231–43.

[2] Fong, Geoffrey T., David H. Krantz and Richard E. Nisbett (1986): 'The effects of statistical training on thinking about everyday problems', *Cognitive Psychology* 18, 253–92.

[3] Kahan, Dan M. et al. (2017): 'Science curiosity and political information processing', *Advances in Political Psychology* 38, 179–99.

[4] Perkins, D. N. (1985): 'Postprimary education has little impact on informal reasoning', *Journal of Educational Psychology* 77, 562–71.

[5] Perkins, David (2019): 'Learning to reason: the influence of instruction, prompts and scaffolding, metacognitive knowledge, and general intelligence on informal reasoning about everyday social and political issues', *Judgment and Decision Making* 14, 624–43.

[6] Kahan, Dan M. (2015): 'Climate-science communication and the *measurement problem*', *Advances in Political Psychology* 36, 1–43.

[7] Kahan, Dan M. et al. (2010): 'Who fears the HPV vaccine, who doesn't, and why? An experimental study of the mechanisms of cultural cognition', *Law and Human Behavior* 34, 501–16.

[8] Kahan, Dan M. et al. (2015): 'Geoengineering and climate change polarization: testing a two-channel model of science communication', *ANNALS of the American Academy of Political and Social Science* 658, 192–222.

[9] Ranehill, Eva et al. (2015): 'Assessing the robustness of power posing: no effect on hormones and risk tolerance in a large sample of men and women', *Psychological Science* 26, 652–6.

[10] Nyhan, Brendan and Jason Reifler (2015): ‘The effect of fact-checking on elites: a field experiment on U.S. state legislators’, *American Journal of Political Science* 59, 628–40.

[11] Henry, Emeric, Ekaterina Zhuravskaya and Sergei Guriev (2022): ‘Checking and sharing alt-facts’, *American Economic Journal: Economic Policy* 14, 55–86.

[12] Pennycook, Gordon et al. (2020): ‘The implied truth effect: attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings’, *Management Science* 66, 4944–57.

Index

Founded in 1893,
UNIVERSITY OF CALIFORNIA PRESS
publishes bold, progressive books and journals on topics in the arts, humanities, social sciences, and natural sciences—with a focus on social justice issues—that inspire thought and action among readers worldwide.

The UC PRESS FOUNDATION
raises funds to uphold the press's vital role as an independent, nonprofit publisher, and receives philanthropic support from a wide range of individuals and institutions—and from committed readers like you. To learn more, visit ucpress.edu/supportus.