SpringerBriefs in Psychology

John Wilcox

Human Judgment

How Accurate Is It, and How Can It Get Better?

Springer

SpringerBriefs in Psychology

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic. Typical topics might include:

- A timely report of state-of-the-art analytical techniques
- A bridge between new research results as published in journal articles and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study or clinical example
- A presentation of core concepts that readers must understand to make independent contributions

SpringerBriefs in Psychology showcase emerging theory, empirical research, and practical application in a wide variety of topics in psychology and related fields. Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, standardized manuscript preparation and formatting guidelines, and expedited production schedules.

John Wilcox

Human Judgment

How Accurate Is It, and How Can It Get Better?

Springer

John Wilcox
Departments of Psychology and Philosophy
Stanford University
Stanford, CA, USA

ISSN 2192-8363 ISSN 2192-8371 (electronic)
SpringerBriefs in Psychology
ISBN 978-3-031-19204-3 ISBN 978-3-031-19205-0 (eBook)
https://doi.org/10.1007/978-3-031-19205-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Testimonies

John Wilcox presents an accessible, insightful take on the questions that should captivate any social scientist: *how accurate is our judgment, and can it be improved?* Wilcox provides compelling evidence about the fallibility of human judgment, dispelling myths and assumptions about expertise, and providing insights on strong indicators of accuracy. The analysis then turns to routes for improving judgment and resultant decision-making, offering a grounded, practical, how-to-guide for individuals and organizations alike. A must-read for those interested in all disciplines, including those of us concerned with improving politics and policy.

– *Rachel George, Ph.D, Lecturing Fellow, at Duke University.*

I just saw a patient with frequent urination and some burning – "You have a urinary tract infection", I told her. Doctors say these things with certainty, and that is where things get interesting. Indeed most of us believe we ARE sure about our diagnoses, but research suggests we shouldn't be!

John Wilcox presents a remarkable new analysis of accuracy in decision-making, how we should understand the uncertainty that surrounds our judgments, and what it means to be well calibrated – where your sense of certainty varies along with the odds your answer is correct. It is good news to me that we doctors are not alone: Judgments are fallible, and it's not just medicine – so are legal decisions, military decisions, and business decisions. This is a fresh look at intuition, heuristics, and calibration in decision-making, starting with grounding in the foundations of these issues in psychology and philosophy, and concluding with real-world examples that make the points relevant and practical. This is a great and educational read – we ALL make decisions every day of our lives. Wilcox is our expert guide on how each of us, and society at large, could do a better job with these, and improve our calibration in the process.

– *Mark Graber, M.D, Fellow of the American College of Physicians and the Founder of the Society for Improving Diagnosis in Medicine, USA*

John Wilcox's book is a timely and lucid study of what it means to think rationally and why it's important for our understanding of human minds.

– *Samuel Gershman, Ph.D., Professor of psychology at Harvard University and the author of "What Makes Us Smart"*

Acknowledgments

I should thank many people for their help in producing this book. Foremost among them are my former advisors at Stanford University, especially epistemologist Ray Briggs and cognitive scientist James McClelland. Both of them have provided large amounts of time and guidance in nourishing various aspects of my academic career, including work related to this book. For that, I am profoundly grateful. Aside from that, I also have a debt of gratitude to Mark Graber, Dan Sperber, James Gross, Mike Frank, Krista Lawlor, Sam Gershman, Mikkel Gerken, Megan Salole, Leïla Mörch, Alan Wilcox, and others who are too numerous to thank here. I appreciate all of their feedback.

However, I am especially grateful to Daniel Kahneman and the late Amos Tversky. I have never interacted with either of them, but their work has had a powerful impact on cognitive psychology, on many other fields, and on my own work. I sometimes think of myself as somewhat following in their intellectual footsteps. For that reason, I dedicate this book to them and to their enduring influence on how we understand human judgment.

Contents

Chapter 1
Introduction

1.1 Judgmental Accuracy: Why It Is Important

Let me tell you a story—an entirely true one. In a relatively quiet city in Australia, a mother had become concerned about her son's mental health. Over time, the son's behavior got worse, and he started having hallucinations and strange beliefs. So he went in and out of hospitals on dozens of occasions. Then, one day, he presented to a police station without a shirt, sweating profusely and claiming that his neighbor was a serial killer. He was sent to hospital that night. Despite this, the son appeared calm the next morning, and so he was discharged by some doctors who reportedly judged that there was "no risk of harm"—notwithstanding the concerns expressed by his mother. So he went home, and, the next day, he heard commands from the gods to kill his family, and so he stabbed them to death—his mother, his younger brother, and his younger sister—146 times with a kitchen knife (Ramsey, 2021; Rennie, 2021).

That story did not have a happy ending, yet I shared it—extreme as it is—because it illustrates an important moral: doctors require accurate judgments, including accurate judgments about risks. Of course, doctors might use different words to describe their judgments, words like "diagnoses," "prognoses," "assessments," "evaluations," and the like. But they certainly require accurate judgments nonetheless, and the quality of what they do depends on it, often in ways that are *very, very* important.

Of course, you might have doubts about that example: you might think that medical errors like this are rare, that the doctors reasoned perfectly well, and that they just got unlucky about a difficult case.

Maybe that is the case, but there are experts who might disagree. While he does not go into details, one forensic psychiatrist reviewed the incident and stated that he was surprised the son's case had not "been taken more seriously" by the doctors (Ramsey, 2021). The courtroom judge also concluded that the son had never received adequate treatment and was failed by the community.

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0_1

But regardless of what we think about this case, later on, we will see study after study indicating that harmful misdiagnoses are not rare either, even for more common conditions like cancer and heart failure. What's more, some medical researchers estimated that as many as 40,000 to 80,000 adults in the United States die each year because of "preventable" diagnostic errors. If those researchers are right, then judgmental inaccuracy exacts a death toll on Americans that is at least 13 times higher than the September 11th attacks—and in just one year alone.

But in any case, whatever we think of this story, arguably the main moral still holds: doctors need accurate judgments, and the quality of what they do depends on it.

Yet a similar moral applies to all of us: to jury members, to the President of the United States—and arguably even to people like you and me!

Jury members need to make accurate judgments about, say, whether a defendant is guilty of some crime. And again, they might use different words to describe their judgments, words like "verdict," "conviction," "decision," and the like. But again, they certainly require accurate judgments nonetheless, and the quality of what they do depends on it. If they have accurate judgments, justice is served, and criminals are sent to prison, but if they do not, criminals live freely with impunity, while innocent people are sent to jail—sometimes for years. And as we shall see, some scholars estimate that at least 4.1% of people who are sentenced to death in the United States are actually innocent. What's more, while some of these innocent people will be exonerated, most of them will be either executed or instead sent to life in prison and then forgotten—all because humans inaccurately judged them to be guilty of horrific crimes that they never actually committed.

The President of the United States also needs to make accurate judgments about, say, the threats posed by a foreign country. Here, accurate judgments can make the difference between a judicious intervention on the one hand and a war waged on false premises on the other—the worst examples of which could needlessly send thousands of sons, daughters, fathers, and siblings to an early death.

Of course, these examples show that judgmental accuracy is not an abstract philosophical or psychological topic: it is extremely important—so much so that it can literally be the difference between life and death, sometimes for thousands, if not millions.

But judgmental accuracy is also important in more mundane contexts. Consider your most important life decisions: examples may include which career to pursue, who to marry, and whether to have a child. Good decisions about each of these might require accurate judgments about a range of topics, such as which career will make you happier, whether your significant other has qualities made of marriage-material, and whether you are well placed to have a child. And when we think about these topics, we can also see how accurate judgments can again distinguish good decisions from poor ones: they can be the difference between a fulfilling career or a depressing one, a successful marriage or a brutal divorce, and so on and so forth.

1.2 The Focus of This Book

So, we all require accurate judgments, but this book offers two pieces of news—one good and the other bad.

The bad news is that while we sometimes have accurate judgments, we often do not, and we do not even realize it—at least in many contexts where we might have thought or hoped otherwise. That is the first theme of this book covered in Chaps. 3, 4, 5 and 6.

The good news, however, is that our judgments can become more accurate, and I hope to share some research with you about how this is so. That is the second, more optimistic and perhaps more important theme of this book—the theme covered in Chaps. 7 and 8.

So this book is about the accuracy of our judgments—about the science of how accurate we are and how to get more accurate.

That said, a caveat is in order: this book's focus is restricted exclusively to judgments about so-called *descriptive* facts—facts about, for example, whether the Democrats will win the next US presidential election, whether climate change is occurring, or whether a defendant killed the victim.

I do not discuss *ethical, moral*, *or evaluative facts* about, for example, whether it is *good* or *bad* if the Democrats win the next presidential election, whether we are *ethically obligated* to decelerate climate change or whether the defendant was *ethically justified* when they killed the victim. These ethical judgments are important, but they are not the focus of the book, in part because questions about their accuracy are more complicated and deserve a book-length treatment in their own right.

From this, it should also be clear that this book is also not about "judgment" in the sense of assigning praise or blame. In the cognitive science literature, the term "judgment" is often used in a way that is entirely different to how everyday people talk about, for example, someone "judging" another person because they did something wrong, or because they wear unfashionable clothing, or what have you.

This book, then, is not about ethical judgments of good or bad, of right or wrong, of blame or praise.

It is just about judgments about descriptive facts, like those listed above.

That said, this book is not entirely irrelevant to ethical judgment and decision-making either. We often need accurate judgments about descriptive facts to inform our judgments about ethical issues: for example, the ethical judgment that we are obligated to decelerate climate change depends on the descriptive judgment that climate change is indeed occurring in the first place. Ethical judgments are compromised if they rely on inaccurate descriptive judgments. For that reason, improving the accuracy of descriptive judgments can also improve our ethical judgments.

Yet the focus is nearly always on questions about the accuracy of our descriptive judgments, as well as how to improve such accuracy.

My hope is to synthesize a range of insights to help us answer these questions. If successful, the payoff should be obvious: we can apply these insights to improve the quality of our decision-making, our organizations, our society, and ultimately our lives. We can then avoid many of the pervasive pitfalls that afflict our judgments and undermine our decisions—sometimes unknowingly, and sometimes with severe or even deadly consequences.

Throughout this book, we will draw on insights from a variety of literatures, including those about medicine, politics, or law—to name but a few examples.

However, two fields are especially important. One is epistemology, the field that involves—among other things—the philosophical study of how humans *ideally would* think. The other is cognitive science, the field that involves—among other things—the scientific study of how humans *actually do* think.

I hope to take you on a tour through both of these fields, drawing on my training as both an epistemologist and a cognitive scientist.

And in the final chapters, we will explore a fusion of the two fields which I call *empirical epistemology* or *normative cognitive science*: the scientific study of how humans *ideally should* think based on empirical evidence about what ways of thinking *actually do* lead to more accurate judgments in real-world contexts.

A word of warning though: the literature is depressing at times, albeit inspiring at others. But hopefully, the payoff for perseverance is worth it: by understanding the infirmities of our judgment, how to avoid them, and how to improve our judgment, this can ultimately help us make better decisions and live better lives—both as individuals and as a collective.

1.3 How This Book Was Researched

A few more words about how this book was researched and written.

Our topic here—the accuracy of human judgment and how to improve it—is vast and pervasive, touching on any area of society where humans need accurate judgments.

As it turns out, though, these areas are *many* areas, and as a result, insights about human judgment are scattered across various fields.

But because these insights are so scattered, this book's literature review was not conducted in a typically systematic but narrow fashion: there is no single database search which will yield insights about human accuracy in fields as diverse as medicine, political judgment, and epistemology, for instance.

Consequently, this book haphazardly but cautiously aims to integrate insights in ways that transcend a narrowly defined literature review. This distinguishes it from typical psychological reviews, since it also enables it to draw on a much broader body of literature than is usual. But that said, lovers of narrow reviews and meta-analyses will find them in the references, since many of them are cited throughout this book.

Regardless, the purpose of this book's eclectic synthesis is then twofold. The first is to help us get a more holistic view of human cognition in general and of judgmental accuracy in particular. The second is to enable researchers concerned with judgmental accuracy in one field to benefit from the insights in another. This latter aim is especially since apropos: it sometimes seems that researchers in different fields are concerned with essentially the same subject matter—the accuracy of human judgment—but it is not obvious that they are talking to each other when they could be.

1.4 Intended Audiences and How to Read This Book

Having shared some thoughts about how this book was written, let me share a few about how it can be read.

For a start, the potential readership of this book is diverse—potentially encompassing academics, organizations, and the broader population at large.

For that reason, some content is tailored more to some groups rather than others. Some epistemologists would be uninterested in my discussion of popular objections to the correspondence theory of truth, for instance, while the general public may very well be uninterested in the epistemological and psychometric discussion of less valid measures of judgmental accuracy.

How, then, do I address this problem—how can I write for multiple audiences with seemingly incompatible interests? My solution is simple: write for all of them but tell the reader to skip the parts they find uninteresting. Maybe that is not the best solution, but in any case, the point is that there is no need to read all of this book. As a reader, just take what you need. That way, hopefully content can be written for different audiences without inflicting torturous boredom on them all.

Finally, I should also mention that I have aimed to write handy chapter summaries for those short on time.

1.5 Structure of This Book

The structure of this book is then as follows.

The next and second chapter of this book focuses on the concept and measurement of judgmental accuracy: what is judgmental accuracy, and how can we measure it? It defines judgmental accuracy as a matter of having appropriate levels of confidence that propositions or ideas are true. Truth is understood as per the correspondence theory of truth: a proposition or idea is true to the extent that it corresponds to reality. The correspondence theory implies that propositions can be objectively true—that is, true in a sense that is independent of people's judgments about the truth. The chapter defends the notion of objective truth against various objections, including the objections that people disagree about truth, that it is sometimes difficult or impossible to determine the truth, and that truths about morality or other topics are subjective. Then, the chapter considers the idea that confidence can be modelled with probabilities: for example, you might be 80% confident that you will stay in your job a year from now, or 99.9999% confident that the sun will rise tomorrow. The accuracy of these confidence levels can be measured in various ways, such as measures of calibration, resolution, and Brier scores (measures which I shall explain then). It finally argues for the superiority of particular calibration and resolution measures of accuracy compared to other measures of accuracy.

Chapter 3 reviews the accuracy of human judgment across various fields, including cross-cultural psychology, medicine, political judgment, and criminal convictions in law. It argues that the available evidence suggests humans are much more inaccurate than they generally would hope or expect, at least in many contexts. It also argues that this has real-world consequences, such as misdiagnoses or false death sentence convictions, both of which result in the needless loss of life. That said, the chapter tempers its claims by arguing that the accuracy of human judgment depends on the context: humans are accurate in some contexts, they are worryingly inaccurate in others, and then they have levels of accuracy that are simply unknown in many more contexts. This is called the *context-dependent model* of human accuracy.

We might then ask what explains why we are so inaccurate. This question is spoken to by the following three chapters of this book: one on so-called *metacognition* (how we think about our thinking), another on the so-called *heuristics* by which people often arrive at their judgments, and another chapter on the *evolution* of our psychology.

Chapter 4 examines how we evaluate our own thinking. It presents evidence of so-called *metacognitive inaccuracy*—that is, inaccuracy in assessing the accuracy of our own judgments. This evidence includes studies that report that, for example, students consistently overestimate their own performance in exams. It then reviews various explanations for the development or persistence of metacognitive inaccuracy: 1) that metacognitive inaccuracy arises because the skills that are necessary for accurate judgments about one's accuracy are the same skills that are lacking and whose absence results in inaccurate judgments in the first place, 2) that people lack the motivational incentives to be metacognitively accurate, 3) that people fail to track their past accuracy, and 4) that people have motivational biases *not* to realize their own inaccuracy. Evidence exists for some of these explanations (such as the lack of record keeping about past accuracy), but not for others (such as the lack of incentives).

Chapter 5 discusses the rationality of the reasoning processes by which we arrive at our judgments. It first distinguishes two conceptions of rationality: epistemic rationality, which concerns the accuracy of how we arrive at our judgments, and pragmatic rationality, which concerns whether our actions conduce to our welfare or the satisfaction of our desires or needs. It then outlines dual-process theory, according to which people arrive at their judgments through one of two kinds of processes: Type 1 processes, which are fast, intuitive, and often less reliable, and Type 2 processes, which are slower, more deliberate, and often more reliable. Then, the chapter discusses the role of motivation and confirmation bias in how we arrive at our judgments. It also discusses some of the most well-known heuristics which explain how we arrive at our judgments, including the representativeness, availability, and anchoring heuristics. Finally, it outlines some pernicious social sources of influence on our judgments, including conformity and industrial influence. Throughout this chapter, we will see how a variety of processes, heuristics, and sources can cause us to arrive at inaccurate or epistemically irrational judgments.

Chapter 6 turns to consider some evolutionary explanations of the function of reason, explanations that each say different things about the inaccuracy of human judgment. One explanation is the intellectualist explanation, according to which reason serves the function of helping humans to arrive at true judgments and to make good decisions. The intellectualist explanation would presumably explain the inaccuracy of human judgment as a by-product of the fact that evolutionary mechanisms are sub-optimal: a biological feature may serve a function which confers an evolutionary advantage on a species, even if the feature does so imperfectly. On this view, the imperfection of human judgment is attributable to the imperfection of evolutionary forces. The other competing explanation is articulated by cognitive scientists Hugo Mercier and Dan Sperber: reason serves the social function of producing and exchanging arguments and justifications. On this account, reason is lazy and biased in producing reasons—but in ways that are evolutionarily sensible in light of its functions—while reason is also demanding and unbiased in evaluating reasons from others. While I endorse neither explanation in this chapter, I present some critical issues for Mercier and Sperber's arguments, and I caution against the harm of confirmation bias.

Chapter 7 introduces the concept of *empirical epistemology*—the study of how humans should think based on scientific studies about what actually improves accuracy in real-world contexts. It describes some origins of empirical epistemology, origins which lie at least partly in work funded by the US intelligence community. It defends the presupposition that there are domain general methods of improving accuracy and that insights in a domain like geopolitics can tell us something about improving accuracy in other domains like medicine. It then outlines variables that correlate with improved accuracy. These include the following: *situational variables* about the environment in which one makes inferences, *motivational variables* about one's motivation, *cognitive variables* about how one seeks out information and draws inferences from it, and *metacognitive variables* that concern how one assesses their own cognition. It also discusses *negative* lessons from empirical epistemology: that is, insights about what does *not* conduce to accuracy. The result is a wealth of insight about variables that can predict and improve judgmental accuracy.

Finally, Chap. 8 provides three categories of recommendations based on these insights: recommendations for improving our judgment as individuals, recommendations for estimating the accuracy of others, and recommendations for conducting our organizations in ways that conduce to success.

While many studies paint a less-than-ideal view of human judgment, my aspiration is that these recommendations can help us move closer to the ideal—that is, closer to accurate judgments. As a result, hopefully we can apply the insights from empirical epistemology in ways that improve our judgments, our decision-making, and ultimately our lives—both as individuals and as a collective.

References

Ramsey, M. (2021, February 12). *WA family murder trial shown CCTV footage*. The Standard. https://www.standard.net.au/story/7125016/wa-family-murder-trial-shown-cctv-footage/

Rennie, L. (2021). *Ellenbrook killings: Man who stabbed mum, siblings 146 times found not guilty of murder*. https://www.9news.com.au/national/ellenbrook-killings-man-who-stabbed-mum-siblings-found-not-guilty-of-murder/89514226-2d72-4241-8514-6512bd9a5c7e

Chapter 2
What Is Judgmental Accuracy: Concepts and Measurement

2.1 What Judgmental Accuracy Involves: Correspondence and Confidence

This book is an exploration of human judgment—an exploration of how accurate it is and how to improve our accuracy. A good starting point in our journey is to ask the question "What is judgmental accuracy? What does it *mean* for humans to be accurate or inaccurate?"

This book's answer is somewhat simple: here, we can think of judgmental accuracy as a concept that in turns involves two further concepts—namely, correspondence and confidence.

Let us consider both concepts first, and then we will look at some concrete measures of judgmental accuracy.

2.1.1 Objective Truth and the Correspondence Theory

At the core of judgmental accuracy is the concept of truth: a judgment's accuracy has to do with how true it is, or how close to the truth it is, and so on.

But what is truth?

"Correspondence" is the dominating answer, at least in analytic epistemology—the academic field specializing in the nature of truth and related topics.

The correspondence theory of truth says that truth is about a relationship between two things. Different correspondence theories can differ over the details about precisely what these two things are, as well as what the relationship between them is (Marian, 2022).

Here, though, we can describe the correspondence theory as a theory about two particular things: propositions on the one hand and reality on the other.

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0_2

For simplicity's sake, we can think of a proposition as being more or less like an idea, such as the idea that Barack Obama was the President of the United States or the idea that there is a snake in your lounge.

Sentences and other things can then express various propositions.

Two different types of sentences can express the same proposition. For example, suppose I say "There is a snake in *your* lounge," and you say "There is a snake in *my* lounge". Both of these are different types of sentences, but they express the same proposition or idea—namely, that there is a snake in your lounge.

Two different propositions can also be expressed by the same type of sentence. For example, suppose I say "I am happy" and you say "I am happy". Both sentences are the same *type* of sentence insofar as they both say "I am happy." But one sentence expresses the idea that *I*—the author—am happy, while the other expresses the distinct idea that *you*—the reader—are happy.

The correspondence theory then specifies what it takes for a proposition to be true. In particular, the theory says a proposition or idea is true just in case it corresponds to way things are in reality. Then, the proposition that there is a snake in your lounge would be true just in case it corresponds to reality—just in case there actually is a snake in your lounge. And it would be false if it fails to correspond to reality—just in case there is no snake in your lounge.

A judgment that a particular proposition is true would itself be true just in case that particular proposition is true, and anything expressing a particular proposition is true just in case the particular proposition is true.

A useful analogy for thinking about the correspondence theory of truth is that of a map and some terrain (Korzybski, 1933). Our propositions—or our judgments about them—are like mental maps which may represent the way reality is. A map is a true representation of reality to the extent that it adequately corresponds to how the terrain really is: for example, if a map says Palo Alto city is south of San Francisco, then this is true just in case Palo Alto actually is south of San Francisco. Similarly, propositions and our judgments about them are true to the extent that they correspond to how reality really is.

Importantly, like maps, propositions or judgments may only correspond to reality to a *particular degree*.

This is similar to how a map does not perfectly represent the terrain. Maps are often imprecise; for instance, they do not perfectly represent the terrain down to the detail of every blade of grass on a field. And sometimes they are somewhat false or inaccurate; a map may become outdated when a new roundabout is installed at an intersection, for instance. Nevertheless, despite such imprecision and falsity, a map may nevertheless correspond to reality to some degree.

Similarly, despite some imprecision and falsity, our judgments may nevertheless correspond to reality to some degree.

To take a historical example, Newton's laws of gravitation were approximately true to the extent that they roughly described how the planets orbited the sun and so forth. But while they were approximately true, they contained some degree of falsity: Newtonian mechanics failed to predict the exact orbit of Mercury, and it was only when relativity theory was proposed that we had a theory which accurately described Mercury's orbit and corresponded to the truth to an even greater degree (Hanson, 1962).

So the correspondence theory is an account of the concept of truth—of what it *means* for something to be true.

It is a simple concept, but it is plausible—and historically prominent. In fact, philosopher Franca D'agostini goes so far as to say that the "[c]orrespondence theory has dominated, almost uncontroversially, the entire history of philosophy" (d'Agostini, 2019, p. 272). Its history reaches also as far back as Plato and Aristotle, with Aristotle claiming "To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, and of what is not that it is not, is true" (Metaphysics 1011b25, quoted in Marian, 2022). A strong argument for the correspondence theory as an account of the meaning of truth is also that it explains why many judgments are intuitively true, including those about the roundness of the earth, evolutionary theory, and many other theories: they are true in virtue of their correspondence to reality.

Importantly, though, the correspondence theory implies that there can be so-called *objective* truths—at least if we define "objectivity" in a particular way. More specifically, some philosophers think that something is objective just in case it does not depend on human judgments for its existence (Enoch, 2011; Railton, 1986). Applying this to "truth," we can say a truth is objective just in case it does not depend on human judgments for it to be true, at least not in the sense that judging that thing to be true by itself makes that thing true. For example, consider the Copernican heliocentric theory—the proposition that the earth revolves around the sun. This proposition would be an objective truth since its truth does not depend on human judgments in this particular sense: it would still be true that the earth revolves around the sun regardless of anyone's judgment about whether it does—regardless of, say, whether you judged that it did revolve around the sun or not. After all, judging the heliocentric theory to be true does not make it true that the earth revolves around the sun. What makes the theory true is that, in reality, the earth actually does revolve around the sun; there is a correspondence between the theory and reality, between the map and the terrain, so to speak. So, the Copernican theory is an example of an objective truth because it is a *judgment-independent truth*—it does not depend on human judgments for its truth in this particular sense.

Contrast this to money. It may be true that a particular monetary note has a value of $10, but it would be true only because humans collectively *judge* that it is true; if everyone collectively agreed that a note had no monetary value, then it certainly would not. So some might argue (rightly or wrongly) that judging something to have monetary value by itself makes that judgment true (at least when enough people have that judgment).

2.1.2 Misconceptions About Objective Truth

Of course, often people—especially outside of epistemology—have objections to the notion of "an objective truth". Perhaps most people have heard one or another person say "There is no objective truth," or something similar.

However, these objections frequently rest on misconceptions, even though they often contain (somewhat ironically) a grain of truth.

I will critically survey these objections here, trying to emphasize both what I think they get right as well as what they get wrong. (However, fans of objectivity may wish to skip this section.)

A disclaimer though: my responses to these objections are not entirely novel or original. Based on my experience, I surmise that these objections frequently come up in introductory courses in analytic epistemology, philosophy of science and critical thinking, and that analytic philosophers frequently have the same responses. Consequently, if there is any merit to these responses, I do not claim them for myself. That said, while I have heard these responses from philosophers, I have never heard these objections from philosophers, at least not in the analytic tradition: so far as I can tell, they come entirely from people who are outside of analytic epistemology—not specialists in it.

Let us now consider these objections.

2.1.2.1 Objection #1: The Diversity of Opinions About Truth

One objection is that there is no objective truth because we have different opinions about the truth and not everyone agrees.

An issue with this objection is that it conflates *truth* on the one hand with people's *judgments* about the truth. For example, there was a time when people disagreed about whether the earth was round, whether it revolved around the sun, and a myriad other things we now know are true. However, the diversity of opinions and lack of agreement does not entail there is no objective truth. At most, it entails merely that if there is an objective truth, we may have different opinions or judgments about it.

That said, the grain of truth to the objection is that even though some objective truths exist and do not differ from person to person, our *judgments* about the truth often do differ from person to person. Truth can be objective, even if our judgments are often not objectively true. This is especially the case since if people disagree with each other and if not all of them can be right, then some of them must be wrong. Put simply, disagreement often entails that if there is an objective truth, at least some people have not found it. Sometimes, disagreement can cause us to reasonably doubt whether we have found some objective truth, but not whether there is an objective truth.

2.1.2.2 Objection #2: The Subjectivity of Truths About Money or Other Topics

A second objection is that there is no objective truth because some truths are not objective, like truths about taste, money, or morality.

The grain of truth to this objection is that some truths arguably are not objective, like truths about money. (That said, morality and other topics are more debatable.)

However, this objection shows that, at best, only *some* truths are not objective, but it does not show that *all* truths are not objective. Even if one supposes that truths about taste, money, or morality are never objective, it does not mean other truths are not—including truths about the roundness of the earth, whether it revolves around the sun, and so forth.

2.1.2.3 Objection #3: Track Record of Failures to Grasp Truth

A third objection is that there is no objective truth because we have been mistaken in the past. After all, many humans once thought that the world was flat, that the earth was a few thousand years old, that the continents did not draft apart, and so forth. In these and many other cases, our judgments about reality were mistaken. How, then, can we be so confident that there are objective truths?—or so the objection goes.

There are three problems with this objection.

First, even if we have been mistaken about reality in the past, this would not imply there are no objective truths. Instead, it would imply that we do not always correctly *believe* in objective truths—we do not always make correct judgments about what the objective truths are. But there may still be objective truths, even if we do not believe in them or have been mistaken about them in the past; it may be an objective truth that the earth is round even if everyone judged to the contrary, for instance.

A second problem is that it arguably only makes sense to say we were mistaken about these topics if we already presuppose there is an objective standard of truth. Otherwise, if there is no objective truth, then what is the standard by which we can confidently assess that our past judgments were mistaken? How could we say we were mistaken about the earth being flat unless we thought it was objectively true that the earth was *not* flat?

The third problem with this objection is that it appeals to a biased sample consisting entirely of humanity's failures. It is true that humanity was mistaken about many topics, but there are many topics where humanity was not. Humanity has for a long time correctly judged that most humans walk on two legs, have a nose, and need food and water—and to breathe—in order to survive. Presumably we for a long time correctly judged that jumping off large cliffs or into volcanoes was dangerous—if not fatal. When we consider a broader sample, we will see that while there are many failures, there are also many, many success stories. Indeed, plausibly humanity depends on success stories like these for their very survival, or else we might be unwittingly jumping off cliffs while being incapable of grasping objective truths about how dangerous it is (and this is a theme we shall return to in Chap. 6). And yet these same success stories also suggest that in at least some cases, there are reasons to think not only that there are objective truths, but that we can also know what they are.

So the objection that there is no objective truth because we have been mistaken in the past is itself mistaken, both because we have often *not* been mistaken and because objective truths can exist even if we have been mistaken.

2.1.2.4 Objection #4: There Is No Way to Tell Truth

A fourth and related objection is that there is no objective truth because we have no way to tell whether there is an objective truth.

The problem with this objection is that it does not distinguish between two very different things: whether there is an objective truth about some things and whether we can know what those objective truths are. There might be an objective truth, even if we do not know what it is.

Second, it again neglects a broader sample of success stories where we seem to be capable of grasping objective truths, like those mentioned above.

Of course, this is not to deny that it is sometimes difficult or even impossible to determine what the objective truth is about some questions; sometimes the evidence is scarce or difficult to interpret correctly. But surely it is premature to outright dismiss the possibility of being able to know *any* objective truths.

Plausibly, the correct attitude varies on a case-by-case basis: sometimes, we can be confident in what the objective truth is, at least to a high degree, while in other cases, we may not be able to be so confident. It all depends on the relevant evidence —if there is any—in a particular context.

And as we shall see in Chap. 7, some people—and some ways of thinking—are much better at figuring out the truth than others. So even in cases where it may be difficult tell truth, there may be a path to find it if we think in the right ways.

2.1.2.5 Objection #5: Truth Depends on Language

A fifth objection is that there is no objective truth because truth depends on words. For example, suppose there actually is a snake in my lounge; then one might say it is objectively true that “There is a snake in John’s lounge,” but the sentence is only true because of how we define words. If the word “snake” meant something more like a unicorn, then it would never be true that “There is a snake in John’s lounge.” In this way, one might object that truth depends on judgments about what words mean, and so truth cannot be objective in the sense that it is independent of human judgments.

The problem with this objection is that it conflates a very subtle but important distinction: it conflates the *truth of propositions* with the truth of sentences that *express propositions*. Even though the latter might depend on judgments, the former does not. Let us consider how this is so.

Recall that two different sentences can express the same proposition (I say “There is a snake in your lounge,” and you say “There is a snake in my lounge”) and two different propositions can be expressed by the same sentence (I say “I am happy” and you say “I am happy”).

What this shows is that propositions and sentences expressing them are not the same thing: two propositions can be expressed by the same type of sentence, while two types of sentences can express the same proposition. It is important to understand that propositions—which we can think of as being like ideas—are different from sentences that express propositions. It is like the difference between a person and a microphone they are talking through. A person may be communicated through a microphone in a similar way to how a proposition or idea may be communicated through a sentence, but the proposition is different from the sentence, similarly to how the person is different from the microphone.

Now the objection points out correctly that our judgments determine the meaning of sentences. We judge that the word "snake" refers to the object we call a "snake" in a sentence, and these and other judgments give our sentences the meaning that they do. But note that even though judgments determine the meaning of sentences, this does not entail that they determine the truth of propositions.

As a result, if the meaning of words in a sentence changes, then what changes is *not the truth* of the proposition, but rather *which* proposition is expressed by the sentence. If we change the meaning of the word "snake" to mean "unicorn," then the truth value of the sentence "There is a snake in my lounge" would also change, and the sentence would now be false. But this would be only because the sentence now expresses a different proposition—namely, that there is a (in our usual words) unicorn in my lounge. But note that the truth of the propositions did not change: it is still true that there is a (using our language) snake in my lounge—or so we suppose. What changes is merely whether the sentence "There is a snake in my lounge" now expresses this proposition or a different one. To return to our analogy, changing the meaning of a sentence is like giving the microphone to a different person. Changing the meaning of a sentence changes which proposition is communicated through the sentence in the same way that giving the microphone to a different person changes which person is communicated through the microphone.

In this way, the objection does not work because the truth or falsity of the propositions does not depend on language, even though which propositions are expressed by a sentence does.

2.1.2.6 Objection #6: The Ambiguity and Vagueness of Language

A sixth objection is that there is no objective truth because language is ambiguous or vague. For example, two people may both know that the actor Brad Pitt is 180 cm tall, but one person might think he is "tall," while another may not. In that case, it may appear that there is no objective truth about whether Brad Pitt is "tall" or not.

What the example arguably illustrates, however, is that some sentences or expressions contain vagueness, or are simply neither true nor false until they are clarified so that they express clear propositions.

For example, the statement "Brad Pitt is tall" might mean many things: that Pitt is taller than 90% of American men (which would make the sentence false), that Pitt is taller than the average American man (which would make the sentence true), or that most people on earth would see Pitt and think that he is "tall" (which could make the sentence either true or false). So the statement may be ambiguous. Furthermore, some people think that tallness is a so-called "vague predicate", which admits of borderline cases: in such cases, it may be difficult or impossible to tell whether someone is tall or not, just in the same way that there are borderline cases where it is impossible to tell whether someone is bald or not.

The point is that even though it may be unclear whether a particular sentence is true or false, either because it is ambiguous, or because it involves vague predicates, this does not mean there is no objective truth. It may just mean that the sentence is either objectively true or false when disambiguated (in cases of ambiguity) or perhaps just that its truth may be difficult to determine (as might be true for cases of vagueness).

2.1.2.7 Objection #7: Alternative Definitions of Objectivity

There are also many objections to the notion of objective truth that rest on alternative definitions of what an "objective" truth is. Some might say an objective truth is one that we can be certain of, or that we should never change our judgments about, or that implies we always have infallible knowledge of reality.

If one defines objective truths in this way, then there may very well be no objective truths.

But that is not how objectivity is defined here; again, a proposition is objectively true just in case it corresponds to reality in a particular sense where it does not depend on human judgments. If there are any such propositions—propositions which correspond to reality not merely because of our opinions about it—then there are, by this definition at least, objective truths.

2.1.3 Why Does It Matter?

So we have gone to some length to discuss some objections about objective truth. One might rightly wonder, "Why does it matter whether there are objective truths? Why spend so much ink trying to defend the idea of objectivity?"

There are three reasons for it which I will describe, albeit somewhat quickly and roughly.

The first is that, obviously, I think there are good reasons to believe in objective truth and the objections to it are far from compelling, as I allude to above.

The second reason is that this book is pointless if there is no objective truth. As we shall see, the accuracy of human judgment is largely a matter of having judgments or beliefs about propositions that are objectively true—at least to some degree. If there is no objective truth, then there is no real judgmental accuracy, and this book becomes completely worthless. After all, the title of this book is, "Human judgment: how accurate is it, and how can it get better?" If there was no objective truth, you might think that I could as well have opened and ended this book with the statement "It's not at all objectively accurate, and it can't get any better—just because all truth is subjective. But thank you for reading." As it turns out, however, I genuinely think there is an objective truth, at least about many topics, and so a book like this is not pointless.

However, there is a third reason that is, from a societal perspective, perhaps the most practically significant: the notion of objective truth is important for society as a whole. Society *needs* objective truth, especially in times of crisis. What I mean is that society's success often depends on acknowledging that there is an objective truth and successfully attempting to find it.

For example, when the COVID-19 pandemic sent thousands to the grave and many more to overflowing hospitals, many politicians, doctors, and researchers understandably wanted to know the objective truth about various questions: What could be done to contain the deadly disease? What is the safest way of reducing infections? Are vaccines safe? Is it safe to fly on planes? What measures will reduce needless deaths?

Of course, many people might have formed false or inaccurate opinions about these questions, and some people may have not cared about the truth at all, but the point is that at least *many* people *tried* to get as close to the truth as they could.

Yet if there is no objective truth—and if no opinion was more capable of corresponding to reality than any other—then they might as well have asked any child off the street for their opinion about these things, and the child's opinion would have been no more true than any other opinion. But of course, this is not what anyone did, and the reason for this is that we all operate as though there are objective truths—whether this is explicit or not, whether we realize it or not.

So while one might deny that there is an objective truth, they might do so only up until we badly need it or else things might go very, very wrong for us.

2.1.4 Degrees of Confidence

Above, we talked about judgments or beliefs as though they are *binary* concepts: you either have a judgment that something is true, or you do not. This binary concept is sometimes useful, since we often seem to think and talk about judgments as though they are such binary concepts. And in such cases, we may say such a binary judgment or belief is accurate just in case the relevant proposition corresponds to reality. For instance, your judgment that there is a snake in my lounge is accurate just in case there really is a snake in my lounge.

But as we have seen, objections to the possibility of objective truth often invoke the fact that we sometimes fail to grasp the truth. And this is true: human history is littered with a litany of falsehoods which we once believed at one time or another.

This shows that certainty is hard to come by; we are often not justifiably certain about a range of things, including the causes of illnesses, the most effective way of containing a pandemic, or the innocence or guilt of a defendant at a trial—to take only a few examples.

However, even if we cannot know for certain what the truth is, a vast number of the things we have judgments about are things we can be more or less confident about: diagnoses, trial verdicts, policy decisions, and many other judgments often involve states of belief that are less than fully certain.

Even the things we are virtually certain about are often not things we cannot know for sure. One might be near certain that they will not die in a freak accident tomorrow, and indeed, the vast majority of people on earth will not, but there are still people who occasionally do experience extremely improbable fatal accidents. That is why they are called “freak accidents.” Despite this, people generally do not even think of the possibility of a fatal freak accident. They make plans and act as though they are certain they will not die in a freak accident. (Think about it: when was the last time you heard someone genuinely say “Let’s have coffee tomorrow, unless I die in a freak accident”?) People act as though they are virtually certain of things that are at best extremely improbable—but not impossible.

Regardless, since many of our judgments are about things we cannot be certain about, we also want a concept of judgmental accuracy that can intelligibly ask whether these less-than-certain states of confidence are accurate or justified. In academia, *formal epistemology* is one of the sub-disciplines that is perhaps more directly concerned with how to measure the accuracy of less-than-certain states of confidence (although other fields, like statistics and cognitive psychology, study this question too).

In formal epistemology, states of confidence are often represented as numbers between 0 and 1, or 0% and 100% in percentage notation. For example, you might be 0.8 or 80% confident that you will stay in your job a year from now, or 99.9999% confident that the sun will rise tomorrow, or 0% confident that any elephants can fly. We can call these judgments or states of confidence “probabilities.” Throughout this book, I will assume that many of our judgments or mental states can be represented as probabilities like these (although there are also other interpretations of probability which may have value and validity in other contexts).

If we can think of less-than-certain judgments as probabilities, then, we have a question: how do we measure the accuracy of these judgments or probabilities? That is the topic of the next section.

2.2 How Do We Measure Judgmental Accuracy: Calibration, Resolution, and Friends

There are various ways that one might try to measure judgmental accuracy—ways we can call putative “measures” of judgmental accuracy. In this section, I will discuss the general desiderata of these measures—that is, the features that would make a putative measure a *good* measure of judgmental accuracy. I will then assess some candidate measures of judgmental accuracy against these desiderata.

2.2.1 Measurement Validity: Internal and External

We want a measure of judgmental accuracy to be a good one, or—as psychologists say—to be “valid.”

What does it mean for a measure to be “valid?”

The meaning of “measurement validity” has changed over the years and is presented somewhat differently from author to author. However, several recent accounts all agree on one important point: the validity of a measurement is more about the validity of the *interpretation* of the measurement (Coaley, 2009; Michael Furr & Bacharach, 2013; Zumbo, 2006). Put differently, these authors loosely agree that a measure is valid to the extent that the measure is accurately interpreted as a measure of what it is intended to measure; on this picture, it is more precise to say that *interpretations* of measures are either valid or invalid, not so much the measures themselves.

For shorthand, though, we can say that a measure of some construct is valid just in case the measure can be accurately interpreted as a measure of the intended construct. For example, someone’s yearly income may be an invalid measure of their happiness: a person’s yearly income does not necessarily tell us much about how happy they are, since rich people can drastically vary in how happy they are. But yearly income is nevertheless a valid measure of how much money they receive each year, arguably by definition.

Anyhow, when it comes to judgmental accuracy, I think measurement validity can be usefully broken down into two further kinds of validity, each of which parallel the distinction between “internal” and “external” validity in experimental psychology and social science (Bryman, 2016; Gleitman et al., 2011).

Loosely speaking, we can say that a measure of judgmental accuracy has *internal* validity to the extent that it measures how accurate one’s judgment is *in the contexts in which it is measured*. For example, we might ask someone to forecast the future about geopolitical topics—such as wars, political elections, and so on. We can then ask them to tell us how accurate they thought their own forecasts were. In this case, their self-evaluation might not be an internally valid measure of their accuracy: their self-evaluation may be biased and may instead be influenced by other things, such as how high or low their self-esteem is or how desperately they want you to believe they are accurate—even if they actually are not very accurate. A better measure of their accuracy would be to look at the track record of those forecasts—to compare what they predict with what actually happens. Surely that would be the more internally valid measure of how accurate those forecasts actually were.

However, a good measure of accuracy will often need something more: *external* validity. Loosely speaking, we can say that a measure of judgmental accuracy has external validity to the extent that, if it is internally valid, it would measure how accurate one’s judgment is in the *other contexts for which it is intended to indicate accuracy*. This is important, since we often want measures of accuracy to generalize beyond the narrow contexts in which accuracy is measured. For example, suppose we want to measure the accuracy of one’s political forecasts to indicate whether they are accurate *in general*. Now their track record of accuracy might be an internally valid measure of their accuracy *on political topics*. But it still might not be an externally valid measure of accuracy *in general*. Perhaps, they might be great at forecasting wars between nations but terrible at forecasting the outcomes of basketball matches, for instance. (That said, I will give reasons in Chap. 7 to think that accuracy in one domain like politics can lead to accuracy in another domain, perhaps including basketball.)

So internal and external measurement validity are then two of the desiderata against which to assess measures of validity. Internal validity is about how good a measure is *in the contexts where measurements are made*; external validity is about how good a measure is in *other* contexts where inferences are made from those measurements. (Of course, there are arguably other desiderata for measures too, like reliability, but these desiderata arguably matter only in terms of their relationship to the other desiderata, like how reliability sometimes indicates internal measurement validity.)

A third desideratum, however, is comprehensibility: ideally, people would be able to comprehend what a measure is saying—to accurately interpret what information the measure is intended to convey. After all, the very purpose of a measure is to provide information to others, information about what the measure is supposed to measure. If the measure is incomprehensible, it is simply not doing its job.

How well, then, do various measures fare against these desiderata?

In response to this question, I have a long answer and a shorter answer.

I explore the long answer in greater detail in a paper that is currently under development and possibly will be published by the time you read this (Wilcox, Work in progress).

But here, for the sake of space, I will only provide the shorter answer.

To do so, I will first articulate what I think is the leading candidate measure of judgmental accuracy, and I will then critique other measures based on how they fare against this leading candidate (although you can skip this critique, if you wish).

A clarification though: In this chapter, although I sometimes use real data, I also critique various measures using hypothetical examples that are somewhat simplistic—and possibly unrealistic. This is because these hypothetical examples illustrate the concerns about validity in a way that I think is simpler and easier to understand. That said, the concerns about validity are not merely hypothetical or unrealistic: to some extent, when I criticize a measure's validity, there is supporting empirical data in the aforementioned paper. There, for example, I report imperfect correlations between my favorite candidate measures on the one hand and some alternative measures on the other, thus reinforcing concerns about the validity of the alternative measures.

2.2.2 A Good Measure of Accuracy: Binned Calibration and Resolution

As mentioned earlier, we often cannot be certain about many things we have judgments about, but we can hope for appropriate degrees of confidence. We can model these degrees of confidence as numerical probability assignments: for example, you might be 80% confident that you will stay in your job a year from now, or 99.9999% confident that the sun will rise tomorrow, or 0% confident that any elephants can fly.

We will then review measures of judgmental accuracy that are about how to measure the accuracy of these degrees of confidence or probability assignments, starting with my preferred candidate.

I call my preferred candidate measure "binned calibration," although it is also sometimes called "calibration-in-the-small" (Lechuga & Wiebe, 2011). Binned calibration takes a set of probability assignments and places them in groups which are called "bins": for example, suppose hypothetically that we take all of your assignments between 0% and 4.9% and place them in one group or bin, we then take all of your assignments between 5% and 14.9% and place them in another bin, and so on for all of your assignments. The result is that we have all of your probability assignments grouped by the categories, even though these categories are somewhat artificial (and I discuss various categorizations more in the aforementioned paper). We can then measure your calibration by seeing how far your probabilities are from the frequency with which the propositions are true. For example, suppose we have a bin containing all of your probability assignments of 90%—or thereabouts. We then consider all the propositions that you assigned probabilities to in that bin: we could suppose these include the propositions that the Golden State Warriors will win the next NBA championship, that you will be in your job a year from now, and so forth. Then if you assign 90% probabilities to propositions that are true only 70% of the time, then you are as they say "miscalibrated" by 20%, at least for that bin of probability assignments.

Let us take a more concrete example to illustrate this using the Good Judgment data. The Good Judgment project was a four-year project funded by the US intelligence community. Among other things, the project aimed to uncover what variables correlated with greater forecasting accuracy, especially about geopolitical topics—that is, topics having to do with the outcomes of wars, elections, international trade agreements, and other global dramas. To do this, Good Judgment recruited thousands of volunteers to make literally millions of forecasts over the four-year period. For example, questions included "Will Bashar al-Assad remain President of Syria through January 31, 2012?" or "Who will win the January 2012 Taiwan Presidential election?" Each forecast was a probability assigned to a proposition in answer to a question. For example, a forecaster might have assigned an 80% probability (or degree of confidence) to the proposition that Bashar al-Assad will remain the President through January 31, 2012. The result is a database of 3,143,460 forecasts which can be analyzed, most of which are forecasts for events which now have known outcomes.

One forecaster from the program was particularly well calibrated: user 5265 (whoever that is). Over the four-year period, this forecaster made a total of 25,685 forecasts. Their binned calibration is depicted in Fig. 2.1 and Table 2.1 (shown on the next page).

As you can see there, this forecaster is remarkably accurate.

Out of all the propositions they assigned approximately 2.5% to, about 2.4% of those propositions are true (although these propositions are not always distinct since, for example, they might assign 1% and 3% to the same proposition at different times). For example, they assigned a probability of approximately 2.5% to the proposition that Egypt would lift the state of emergency in Sinai by April 25, 2015, to the proposition that Iran would release Washington Post correspondent Jason Rezaian before June 10, 2015, and to 7,109 other propositions—approximately 2.4% of which were true!

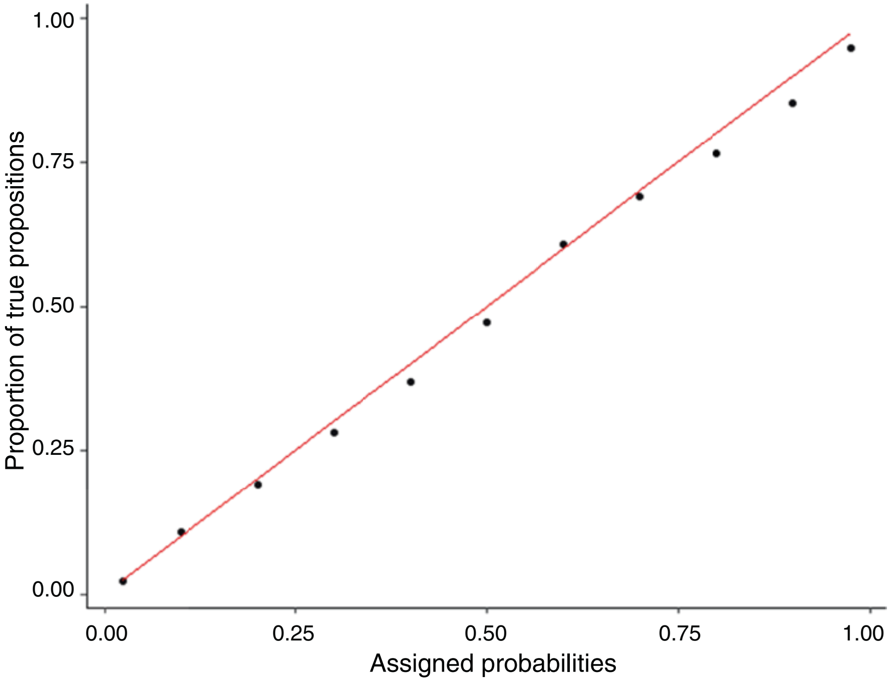

Fig. 2.1 Calibration graph for user 5265

Table 2.1 Calibration table for user 5265

Middle value of the bin	Percentage of true propositions per bin	Number of forecasts per bin
2.5%	2.4%	7,111
10%	10.9%	3,655
20%	19.2%	1,920
30%	28.0%	1,516
40%	36.7%	1,214
50%	47.2%	775
60%	60.7%	1,056
70%	69.1%	1,246
80%	76.6%	1,528
90%	85.3%	2,485
97.5%	94.9%	3,179

They were also pretty well calibrated for other categories: for instance, out of all the propositions they assigned approximately 80% to, 76.6% of those were true—not perfect, but still pretty good. In this case, they assigned a probability of approximately 80% to the proposition that Takeshi Onaga would be elected governor of Okinawa Prefecture in 2014 and to 1,527 other propositions—76.6% of which were true.

The binned miscalibration score for this individual is 1.6%, meaning that their probability forecasts were miscalibrated by only 1.6% on average!

This individual demonstrates that one can be calibrated or accurate when assigning probabilities to *unique events* too. Many of the propositions they made inferences about were quite unique: they concerned specific elections, geopolitical crises, and prisoner situations that happened (or did not happen) once in human history. For example, consider the proposition that Iran would release Washington Post correspondent Jason Rezaian before June 10, 2015. Rezaian was imprisoned only once by Iran and was released only once by Iran (although he was released after June 10, 2015). In this and countless other cases, user 5265 made calibrated inferences about events that are unique, at least at some level of description.

This is significant because some think it is only intelligible to assign probabilities to repeatable events—like the outcomes of coin flips or die rolls—but not to unique events. (This is closely related to the what is known as the "problem of the single case" in philosophy (Hajek, 2019).) User 5265 shows this is not the case: one can assign probabilities to unique events, and they can potentially have superb calibration and accuracy in doing so.

Superb calibration like this is not entirely attributable to luck either. We can tell this for a range of reasons: calibration often improves with practice, calibration is correlated with variables that we expect to correlate with accuracy (such as statistical or probabilistic training), and most people perform far better than simulations of users who randomly assign probabilities and whose accuracy is demonstrably attributed to nothing but chance (Wilcox, Work in progress). These are all things we would not expect if poor or good calibration was entirely due to chance. For that reason, we can tell user 5265 was genuinely well calibrated and not purely by chance.

But not all forecasters were so good, however: one particularly bad example is user 1890. We can see this from Fig. 2.2 and Table 2.2 (shown on the next page).

In this case, they have terrible calibration, at least for some important categories. For example, if you take all the propositions they assigned 97.5% to—that is, all the propositions they reported being nearly certain of—only 62.3% of those propositions are true. In that case, they are miscalibrated by a margin of 35.2%, which is pretty huge! They also reported 100% probabilities for many falsehoods: that the United Nations would pass a new resolution concerning Iran by April 1, 2012; that a Turkish military force would invade or enter Syria between October 9, 2012, and November 30, 2012; that China would seize control of Second Thomas Shoal (a coral reef) by December 31, 2014; and so on. In fact, they assigned 100% probabilities to 169 propositions, but 63 of these were false, about 37% of the propositions they appeared to be certain of! So time after time, it looked as though they were certain about things that—in fact—were false.

Their poor calibration cannot be attributed to small samples sizes either. Sure, they only made a few forecasts for some categories: take their six forecasts which approximated 90%, for instance. But they produced a large sample of forecasts in other categories: take their 252 forecasts approximating 2.5% as an example. (And if you like frequentist confidence intervals for sample proportions, a quick 95% interval analysis indicates that in the 2.5% bin, 25% is a lower bound for true propositions and 37% is an upper bound—far from perfect calibration!)

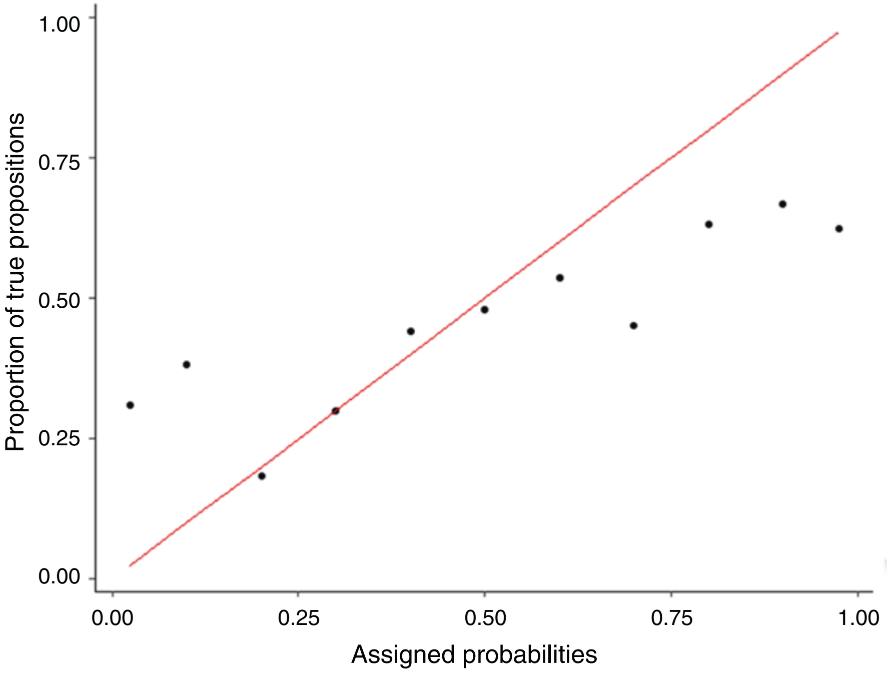

Fig. 2.2 Calibration graph for user 5265

Table 2.2 Calibration table for user 1890

Middle value of the bin	Percentage of true propositions per bin	Number of forecasts per bin
2.5%	31.0%	252
10%	38.1%	21
20%	18.4%	38
30%	29.8%	124
40%	44.0%	25
50%	48.1%	129
60%	53.6%	28
70%	45.2%	31
80%	63.0%	46
90%	66.7%	6
97.5%	62.3%	191

So that is an illustration of binned calibration.

A quick mathematical aside for technically oriented readers (others should skip this paragraph). Binned calibration can be calculated in different ways. Some scholars, for example, present calibration measures which square the distance between a probability on the one hand and the frequency of true propositions on the other (Dunn, 2015). In this case, if your 90% probability assignments hit on the truth only 70% of the time, then your miscalibration score for that category would be $(.9 - .7)^2 = .2^2 = .04$, not .2 or 20%. Such a measure would sometimes deliver different rankings of probability judgments. This is because the squaring procedure could more harshly penalize people who are rarely but severely miscalibrated than people who are frequently but mildly miscalibrated, while the non-squaring procedures could more harshly penalize in the opposite direction. As a result, in some cases, the squaring procedure would rank the rarely but severely miscalibrated individual better than the frequently but mildly miscalibrated individual, but the non-squaring procedure would reverse the rankings for these two individuals. In some cases, the squaring procedure may be better than the other, such as in cases where severe but rare miscalibration should be penalized more harshly. In other cases, the non-squaring procedure is adequate, but it has the added virtue of increased comprehensibility—at least for the less mathematically oriented interpreters of the measure.

In any case, the purpose here is to illustrate what a good measure of accuracy is: binned calibration.

However, like every measure: it has its limitations. In particular, binned calibration requires many probability estimates for a bin before you can have confidence about the accuracy in that bin. If a person is 90% confident in something and that thing turns out to be completely false, maybe their judgments are still accurate overall: after all, even a perfectly calibrated individual will assign 90% probabilities to things which are false 10% of the time, and maybe they just got unlucky this time. To confidently determine whether their 90% confidence levels are accurate, we need to sample *many such* confidence levels—maybe even 50 or more, not just one.

So binned calibration, as a measure of accuracy, requires multiple probability assignments, a *track record*—as I call them in another paper (Wilcox, 2022).

But this, I claim, is an unavoidable limitation: measuring accuracy requires large quantities, at least in many cases. To reiterate, many of our judgments are necessarily probabilistic—they are necessarily less-than-certain. As soon as that is the case, the only way we can confidently measure the accuracy of such judgments is by measuring the accuracy of *many* such judgments. Simply looking at one or a handful of cases is often not enough, merely because we might get lucky or unlucky, the world may conspire for or against us, and it may produce a biased sample of judgments that misleadingly look accurate or inaccurate. That said, sometimes a handful of judgments is enough to get some insight, as I discuss in my long answer to the question of how to measure judgmental accuracy (Wilcox, Work in progress). But other times, it simply is just not enough: to be confident in who is accurate and who is not, large track records of calibration are often necessary (Wilcox, 2022).

However, calibration is not the only thing we may care about; various scholars have also pointed out that *resolution* is desirable too, at least in the presence of calibration (Dunn, 2015; Tetlock & Gardner, 2015). Here, resolution is a matter of having probability assignments that are close to 0% or 100%—a matter of having assignments that are "informative" in this specific sense. After all, one could in principle become calibrated by randomly assigning a probability of 50% to answers in a series of yes-or-no questions, but their assignments would not be particularly helpful or informative. For that reason, it is often appropriate to measure judgmental accuracy with *both* calibration *and* resolution, and I discuss resolution scores and measures of resolution in my longer answer (Wilcox, Work in progress).

That said, the kind of calibration I have discussed here is only one kind: binned calibration at the level of individuals. There are also other measures of calibration and accuracy that we will turn to now.

2.2.3 Less Good Measures of Accuracy

2.2.3.1 Unbinned Calibration

Another measure of calibration is what I call "unbinned" calibration, sometimes also called "calibration-in-the large" (Crowson et al., 2016) or "bias" (Lechuga & Wiebe, 2011).

Unbinned calibration is very simple to calculate. First, you take all of the probability assignments, and you average them. This is what makes it "unbinned" (or we might say "one-binned"): unlike binned calibration, the assignments are not placed in separate bins first. Let us take a small sample of assignments to illustrate this measure (even though good measures of judgmental accuracy should often have many assignments). Suppose you assign four probabilities to four propositions: a probability of 90% to one proposition, 80% to another, 70% to another, and 60% to another. Then, unbinned calibration requires you to average these probabilities together: for example, we get an average of 75% from $\frac{9.+.8+.7+.6}{4}$. Finally, unbinned calibration requires us to subtract the proportion of true propositions from the average probability. So suppose three out of four of the propositions are true. Then, the proportion of true propositions is 75%, and subtracting that from the average probability assignment results in a score of 0%. In this case, then, you have an unbinned miscalibration score of 0%—perfect calibration.

If you have perfect binned calibration and a large enough sample of probability assignments, then you will have perfect unbinned calibration.

The problem with the unbinned measure, however, is that the reverse does not hold: in principle, you could have horrible binned calibration and perfect unbinned calibration. For example, suppose you have 200 probability assignments. One hundred of the assignments assign a probability of 50% to 100 propositions, 75% of which are true. The other 100 assignments assign a probability of 100% to 100 propositions, 75% of which are true. Here, you have a very bad binned calibration score: Your 50% assignments are underconfident by 25%, and your 100% assignments are overconfident by 25%, so you are 100% confident in things which are false 25% of the time! However, you have a perfect unbinned calibration score: if you consider the 200 assignments for the 200 propositions, the average probability assignment is 75%, and the proportion of true propositions is 75%, so the measure says you are perfectly calibrated when, in this case, you are not. Clearly, unbinned calibration is an invalid measure of judgmental accuracy in this context, while binned calibration is superior.

Of course, these are theoretical arguments as to why unbinned calibration is a limited measure of calibration.

However, in my longer answer, I provide empirical results using real-world data which show that unbinned calibration does not correlate especially well with binned calibration, so it is less preferable for that reason (Wilcox, Work in progress).

Nevertheless, unbinned calibration can be useful for detecting the absence of accuracy given a sufficiently large sample of assignments, even if it is not optimal for detecting the presence of accuracy.

2.2.3.2 Brier Scores

Another measure of accuracy is the widely used Brier score (Brier, 1950). The Brier score is calculated as follows. Suppose you consider some proposition *q* and assign a probability value of *P*(*q*), where *P*(*q*) is the probability assigned to *q*: say you attach a probability of 60% to the proposition that Biden will win the 2024 US presidential election, so *P*(*q*) = 0.6 and *q* is the proposition that Biden will win. Then, let *T*(*q*) be the truth value of that proposition, where *T*(*q*) = 1 if *q* is true and *T*(*q*) = 0 if *q* is false. For example, suppose it turns out that it is true that Biden will win the 2024 US presidential election. In that case, *T*(*q*) = 1 and *T*(¬*q*) = 0. Then, the Brier score in general is calculated as follows:

$$Brier = \left(T\left(q\right) - P\left(q\right)\right)^2 + \left(T\left(\neg q\right) - P\left(\neg q\right)\right)^2$$

In our specific example, then, the Brier score for the probability assignment to *q* is as follows:

$$Brier = \left(T\left(q\right) - P\left(q\right)\right)^2 + \left(T\left(\neg q\right) - P\left(\neg q\right)\right)^2 = \left(1 - 0.6\right)^2 + \left(0 - 0.4\right)^2 = 0.32$$

In the literature on geopolitical forecasting, the Brier score is the standard measure of the accuracy of a forecast. The mean Brier score is also the standard measure of the accuracy of a forecaster.

The Brier score is often a useful indicator of judgmental accuracy, so I do not at all condemn its use.

That said, I think binned calibration is occasionally preferable to Brier scores. This is for two reasons.

The first is that Brier scores are less readily comprehensible for many people, at least compared to binned miscalibration scores and resolution scores. For example, a miscalibration score of 0.025 or 2.5% is very readily comprehensible: it means that someone was miscalibrated on average by about 2.5%. It is also obvious that the lower the miscalibration score, the better. Furthermore, a resolution score of 0.9 or 90% is very readily comprehensible: it means that someone's most extreme probability assignment was on average 90%. However, it is more difficult to translate a Brier score of 0.32 into anything meaningful. Of course, it is possible, but one needs more time and familiarity with Brier scores to get a sense of what it means. Furthermore, for some people, it is not immediately intuitive that the lower the Brier score, the better. Usually, people think of scores as things we want to *maximize*—not things we want to *minimize*. It might not be obvious that we want to minimize a "Brier" score at first. But it is obvious that want to minimize a "*mis*calibration" score. In short, Brier scores are less comprehensible and interpretable, at least for many people.

This also leads to the second reason that Brier scores are often less preferrable: it is sometimes unclear whether higher Brier scores are attributable to miscalibration or to low resolution. This is important, because sometimes the evidence is scarce, and any reasonable probability assignment will not be particularly resolute, simply because there is little information available. It is well known, for instance, that geopolitical events are generally more difficult to predict the further they are in the future (Tetlock, 2005; Tetlock & Gardner, 2015). In that case, the Brier score for a forecast would necessarily be relatively high (which, to remind you, is a bad thing, as per the preceding paragraph). When Brier scores are high, however, we might be curious to at least know whether the assignments are calibrated. The Brier score by itself does not give us this information, but the miscalibration score does. Consequently, miscalibration scores can sometimes help us better distinguish bad inferences in the presence of much evidence versus good inferences in the presence of little.

Let us make this more concrete with a very simplified example. Suppose there are two scenarios. In the first scenario, someone—call her "sensible Sarah"—has some informative but less than decisive evidence about 100 propositions. She is perfectly rational and accurate, so she assigns 80% to each of the propositions, and exactly 80% of them turn out to be true. In this case, she has a perfect miscalibration score of 0%; she is perfectly well calibrated! In the second scenario, someone else—call him "radical Mitchell"—has slightly more informative but less than decisive evidence about 100 other propositions. However, radical Mitchell is much less rational than Sarah, so he becomes overconfident and assigns each proposition a probability of 99% when, in fact, only 84% of the propositions are true. In this case, Mitchell is radically overconfident, being nearly certain (i.e., 99% confident) in propositions which are actually false 16% of the time. Instead of being 99% confident in the propositions, we can suppose he should only be 84% confident in them.

However, according to the Brier score, Mitchell does better than Sarah: Mitchell's average Brier score is 0.31, whereas Sarah's average Brier sore is 0.32. The reason for this is partly that although Mitchell is less calibrated than Sarah, he gets rewarded because he is even more resolute when he turns out to be right.

If the Brier score was the only metric by which we assessed the accuracy of Sarah and Mitchell, then Mitchell would look more accurate than Sarah, but clearly this is not the case: Sarah is perfectly calibrated and merely has a little less evidence than Mitchell, while Mitchell is lucky to have a little more evidence but is grossly overconfident and miscalibrated.

This, then, is a concrete example where the Brier score is a less valid measure of accuracy, at least compared to calibration and resolution. Other examples like this could be given as well.

Ultimately, then, I prefer calibration and resolution to Brier scores.

Of course, one putative advantage of Brier scores is that they are "proper" scoring rules. "Proper" here is a technical term—not to be confused with our ordinary understanding of the term "proper." A scoring rule is proper just in case it is always in someone's interests to announce their true probability assignments (Seidenfeld, 1985). This is best understood by contrasting it to an "improper" scoring rule. One such example is the linear score: here, the score of a probability assignment $P(q) = v$ for any proposition q is $|T(q) - v|$. Suppose, for instance, you have 80% confidence in each proposition in a set of propositions. For technical reasons, if each proposition actually had an 80% chance of being true, then you would probably get a better score by this scoring rule if you announced that you were *more* confident in the propositions than you really are. This, however, is not the case for a proper scoring rule like $(T(q) - v)^2$: there would not be any probable advantage in announcing that you are more confident than you really are.

The binned miscalibration scores are also proper: it is always in the agent's interests to announce their true probability assignment (Wilcox, Work in progress).

But, as mentioned, compared to Brier scores, they are more readily comprehensible, and they are sometimes more informative about whether one is less calibrated and more resolute or instead more calibrated and less resolute.

2.2.4 *Measures of Collective Accuracy*

The previous three measures of accuracy—binned calibration, unbinned calibration, and Brier scores—focused on measuring the accuracy of individuals.

However, sometimes researchers are interested in measuring what we might call *collective accuracy*—the accuracy of groups, as in cross-cultural psychology (Lechuga & Wiebe, 2011).

There are various candidate measures of collective accuracy.

2.2.4.1 Unweighted Binned Calibration

One measure is to use binned calibration in a particular way, but at the collective level: that is, to place the probability assignments from the group into the bins and then score them for accuracy. (I call this *unweighted* binned calibration for reasons I will explain shortly.) For example, suppose you, I, and some other people in our group assign 80% probabilities to some propositions: then, each of our probability assignments are placed in the same bin and scored for accuracy as though we were one individual. Perhaps we collectively assigned 80% to 100 propositions, 75% of which were true, and so we have a collective miscalibration score of 5% for that bin.

This measure might be useful in some contexts, but it also has its limitations. In particular, individuals with more or less accuracy might be over or underrepresented in particular bins. For example, some miscalibrated individuals may be overrepresented in some bins, making the rest of the group appear worse than it is. Suppose, for instance, we have a friend in our group, "radical Mitchell," who assigns 100% probabilities to many propositions, all of which are true only 70% of the time. And suppose the rest of us are much more calibrated and much less confident about things, assigning probabilities no greater than 90% to things. Then, radical Mitchell will make us all look bad, since the binned calibration score will say the group is miscalibrated by 30% in the 100% probability bin, even though it is really just Mitchell's fault and no one else made any 100% probability assignments. So collective binned calibration is limited insofar as it can underrepresent or overrepresent individuals in particular bins, making the group look better or worse than it actually is. One might think this could then make the measure of accuracy invalid, at least if it is intended to measure the accuracy of the group in a more representative way.

Since this measure of calibration does not weight people's contributions to the collective scores based on what fraction of the group they (or their assignments) comprise, I call this *unweighted binned calibration*.

2.2.4.2 Unbinned Calibration

Unbinned calibration could be applied to the collective level to measure collective accuracy. You and I might try measure our accuracy as a group, so we average our probability assignments and subtract the proportions of true propositions from that average, for example.

But unbinned calibration has its limitations too. To see this, we can modify the earlier example from unbinned calibration for individuals. Suppose instead the 100 assignments of 50% came from me and the other 100 assignments of 100% came from you. Then, I will be underconfident by 25%, since 75% of the propositions I am 50% confident in are true, while you will be overconfident by 25%, since 75% of the propositions you are 100% confident in are true. But according to our collective unbinned calibration score, we have perfect accuracy: our average probability assignment is 75%, and so is the proportion of true propositions, leaving a remainder of 0% miscalibration when the latter is subtracted from the former. Clearly, this can be an invalid measure of collective accuracy.

2.2.4.3 Brier Scores

Brier scores again can be used at the collective level, but they again inherit the limitations of Brier scores when applied at the individual level: they are less readily comprehensible for some people, and they sometimes fail to distinguish calibrated (but irresolute) assignments from miscalibrated (but more resolute) assignments.

2.2.4.4 Weighted Binned Calibration

I think that, generally, the ideal measure of collective accuracy would be binned calibration but averaged in a particular way. More specifically, binned calibration scores should be calculated for individuals and then the collective score should be formed from these individual scores. For example, if we are in a group of 100 people and radical Mitchell is badly miscalibrated, then his scores should be averaged to only take up his fair share of the group. Maybe everyone has a miscalibration score of 0% when averaged over all the bins, except Mitchell who has a miscalibration score of 30%. Then, the collective miscalibration score averaged over all the bins should be $\frac{.3+0+0+\ldots+0}{100}=.003$ or 0.3%. Likewise, if everyone is perfectly calibrated in the 100% category, except for Mitchell who has a miscalibration score of 30% for just that category, then his collective miscalibration score should be averaged in a similar way to equal 0.3%.

Of course, this requires a lot of probability assignments, but again, this is often an inescapable fact about measuring accuracy in less-than-certain contexts: to determine accuracy, we often need many such assignments.

2.3 Summary

In this chapter, we explored two questions: what is judgmental accuracy, and how can we measure it?

To answer the first question, we explored a couple of concepts.

The first concept is correspondence. More specifically, we explored the *correspondence theory of truth*, according to which *propositions*—or ideas—are true when they correspond to reality. Propositions—or our judgments about them—are then like maps that are true to the extent that they correspond to reality. Like maps, propositions or judgments may correspond to reality to only a degree; like maps, they are sometimes imprecise or contain some falsity, even if they are still somewhat close to the truth. Newtonian mechanics is like this: it described the laws of physics very well up to a degree, although relativity theory is now seen as a more precise description of reality. Furthermore, according to the correspondence theory, propositions can be *objectively true* only if their truth does not depend on human judgments (at least not in the sense that judging something to be true by itself makes that thing true). The proposition that the earth revolves around the sun is like this; it is true, even if, say, no one judged it was true. But one might think money is not like this; it may be true that a note has a value of $10 but only because everyone in the relevant community agrees or judges that it has a value of $10.

There are many objections to the notion of objective truth. These often rely on misconceptions about objective truth. For example, people often say there is no objective truth because not everyone agrees, or because we cannot tell the truth, or because we have been mistaken about truth in the past, or because truths about money and taste are subjective. However, these particular objections conflate whether there is an objective truth with whether everyone agrees about it, or whether we can tell the truth, or whether we have been mistaken about it in the past, or whether other truths are subjective: there may be an objective truth about some things even if we disagree about it, even if we cannot tell the truth, even if we have been mistaken about it and even if there are no objective truths about particular other topics. (There are other objections to the notion of objective truth which were explored in this chapter too.) That said, if objective truth is defined differently—as, say, requiring us to be infallibly knowledgeable about everything—then there may very well be no objective truths on that particular definition. However, this book does not adopt that definition: here, a proposition is objectively true just in case it corresponds to reality and does not depend for its correspondence on the judgments of humans, at least not in the sense that such judgments by themselves make these judgments true.

Having defended objective truth from these objections, I explained why it was important: because there are good reasons to believe in objective truth and no strong reasons to doubt it, because this whole book is predicated on the existence of truth and because society needs to find objective truths, especially when navigating crises like COVID-19.

Aside from correspondence, the second concept that was explored was confidence. Often, we cannot be certain that propositions are true, but only more or less confident. Importantly, we want to be able to talk about the accuracy of these less-than-certain judgments. To do that, we can follow a tradition in formal epistemology, representing these states of confidence with numbers and calling them “probabilities”: for example, you might be 80% confident that you will stay in your job a year from now, or 99.9999% confident that the sun will rise tomorrow, or 0% confident that any elephants can fly.

We can then turn to consider the second question: how do we measure the accuracy of less-then-certain judgments—of these probabilities or degrees of confidence?

To answer this question, we first outlined three desiderata for measures of judgmental accuracy. The first is that they have internal measurement validity: that they are correctly interpreted as measures of accuracy *in the contexts where the measurements are made*. The second is that they have external measurement validity: that they are correctly interpreted as measures of accuracy for *other contexts where they are intended to represent judgmental accuracy*. The third is that they are comprehensible: that people can understand what the measures mean.

I then outlined some measures of accuracy, starting with my preferred candidate and then evaluating the other candidates against this preferred candidate.

This preferred candidate is called *binned calibration*—also known as “calibration-in-the-small”. Binned calibration assigns each probability to a category called a “bin”: all your probability assignments around 90% would be assigned to the bin containing all assignments between 85% and 94.9%, for instance. Then, the accuracy of these probabilities is measured by measuring the difference between the probabilities on the one hand and the frequency with which they attach to true propositions on the other: for instance, perhaps you are around 90% confident in propositions which are true only around 70% of the time, in which case you are inaccurate or “miscalibrated” by around 20% for that bin. Scores across the bins can then be averaged to obtain one’s overall miscalibration score.

A limitation of binned calibration is that it sometimes requires many assignments for a bin before it can accurately measure accuracy, but this is unavoidable for any measure. When it comes to probabilities, the only way to measure accuracy is often with many probabilities. Otherwise, too few probabilities can be useless. For instance, you might be perfectly accurate in your judgments, but also 90% confident in a proposition that later turns out to be false. After all, people sometimes are unlucky, and even a perfectly accurate person will be 90% confident in things that are false 10% of the time. When we try to measure someone's accuracy with too small a sample size, it is often difficult to tell whether they are, say, accurate or simply got lucky.

Yet calibration is not the only thing we care about: it is also important to consider how resolute we are—that is, how close our judgments are to 100% or 0%. We could be perfectly calibrated by randomly assigning 50% probabilities to a series of answers to yes/no questions, for instance, but this would not be helpful since we want answers that are not only calibrated but also informative in the sense of being close to 0% or 100%.

I then compared particular other less desirable measures of accuracy against binned accuracy to show their limitations. This includes unbinned calibration: a measure where all the probability assignments are averaged without placing them in bins and the frequency of truth among the relevant propositions is subtracted from the resulting average. The comparison also included Brier scores, a common measure of accuracy that is often less readily comprehensible and also can fail to distinguish judgments that are less resolute but more calibrated from judgments that are more resolute yet less calibrated. That said, these and other measures can have their utility in some contexts.

I also argued for my preferred measure of *collective accuracy*—that is, the accuracy of a group. This measure involves calculating the binned calibration of individuals and then averaging such calibration scores across the group in the appropriate way. Again, it requires many probabilities to measure this, but this is a necessary cost for greater measurement validity.

That said, although I have reviewed these candidate measures and although I have made my preferences clear, such measures will not always be used throughout this book; sometimes they will be, but sometimes other indicators of accuracy can be useful too. The aim, then, is to form an accurate picture of human judgment by drawing on a diversity of sources of information.

So, how do humans fare against various measures of accuracy?

That is the topic of the next chapter.

References

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.

Coaley, K. (2009). *An introduction to psychological assessment and psychometrics*. SAGE.

Crowson, C. S., Atkinson, E. J., & Therneau, T. M. (2016). Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research, 25*(4), 1692–1706.

d'Agostini, F. (2019). Misunderstandings about truth. *Church, Communication and Culture, 4*(3), 266–286. https://doi.org/10.1080/23753234.2019.1667252

Dunn, J. (2015). Reliability for degrees of belief. *Philosophical Studies, 172*(7), 1929–1952. https://doi.org/10.1007/s11098-014-0380-2

Enoch, D. (2011). *Taking morality seriously: A defense of robust realism*. Oxford University Press.

Gleitman, H., Gross, J. J., & Reisberg, D. (2011). *Psychology* (8th ed.). W. W. Norton & Co.

Hajek, A. (2019). Interpretations of probability. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/

Hanson, N. R. (1962). Leverrier: The zenith and nadir of Newtonian mechanics. *Isis, 53*(3), 359–378.

Korzybski, A. (1933). Science and sanity; an introduction to Non-Aristotelian systems and general semantics. The International Non-Aristotelian Library Publishing Company; The Science Press Printing Company.

Lechuga, J., & Wiebe, J. S. (2011). Culture and probability judgment accuracy: The influence of holistic reasoning. *Journal of Cross-Cultural Psychology, 42*(6), 1054–1065. https://doi.org/10.1177/0022022111407914

Marian, D. (2022). The correspondence theory of truth. In *The Stanford encyclopedia of philosophy* (Summer 2022). https://plato.stanford.edu/archives/sum2022/entries/truth-correspondence/

Michael Furr, R., & Bacharach, V. R. (2013). *Psychometrics: An introduction* (2nd ed.). SAGE.

Railton, P. (1986). Moral realism. *The Philosophical Review, 95*(2), 163–207.

Seidenfeld, T. (1985). Calibration, coherence, and scoring rules. *Philosophy of Science, 52*(2), 274–294. https://doi.org/10.1086/289244

Tetlock, P. (2005). *Expert political judgment: How good is it? How can we know?* Princeton University Press.

Tetlock, P., & Gardner, D. (2015). Superforecasting: The art and science of prediction. *Broadway Books*. https://doi.org/10.1201/b15410-25

Wilcox, J. (2022). *Credences and trustworthiness: A calibrationist account*. (Under Review).

Wilcox, J. (Work in progress). *Measures of accuracy*.

Zumbo, B. (2006). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26).

Chapter 3
What We Think: The Accuracy of Our Judgments

So we have explored some concepts to do with judgmental accuracy which we saw was largely about having appropriate confidence that our ideas are true—that they correspond to reality. We also looked at some concrete measures of accuracy, and I argued that some measures are better than others.

Now we can explore the question: how accurate is human judgment?

The answer I will give is somewhat simple, but it is true: it depends on the context. In some contexts, we know humans are highly accurate. In others, we know humans are worryingly inaccurate. I call this the *context-dependent account* of human accuracy.

These claims are probably not so surprising.

But despite this, people often make full-blown generalizations in either direction, saying things like “Human are awfully inaccurate!” or “Humans are wonderfully accurate!” Yet these statements themselves are to some extent inaccurate: sometimes we are great, and other times we are very worryingly inaccurate.

However, what is newsworthy is the fact that the worryingly inaccurate contexts are much more pervasive than many would hope or expect, as I will argue throughout this chapter.

But that is the only the first newsworthy item.

3.1 Who Is Accurate: How Society Flies Blind

The second newsworthy item is that, in many more contexts, we simply do not know how accurate humans are. This is worrying, especially because of the prevalence of the contexts where we know humans are worryingly inaccurate. Hence, we have little ground for unquestioning optimism. But despite this, inaccuracy is often unmeasured, and relatively few people seem to care.

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0_3

Let us examine this point in more detail though, since it is at once both important and bizarre.

Consider a person making judgments about a topic—any topic. It could be an intelligence analyst judging the likelihood of an attack from a foreign country, or a policy maker deciding whether a policy will be effective, or a doctor diagnosing the illness of a patient, or a jury member deliberating on the guilt of a defendant. What is obvious is that humans are fallible; sometimes, we are accurate and get it right, but sometimes we are inaccurate and get it wrong. What is often not known, however, is how accurate these humans are.

But in all of these contexts, these people are either accurate or not—well calibrated or not.

Perhaps, for example, if the accuracy of their judgments were measured, they would have a perfect calibration graph—just like user 5265 below who we saw in the previous chapter. User 5265 assigns approximately 2.5% probabilities to things that happen 2.4% of the time, and they are calibrated in their other judgments too, as per Fig. 3.1.

So any individual could be well calibrated like user 5265.

Or perhaps they would have an imperfect calibration graph—just like user 1890 who assigns approximately 97.5% probabilities to things that happen only 62.3% of the time, as in Fig. 3.2.

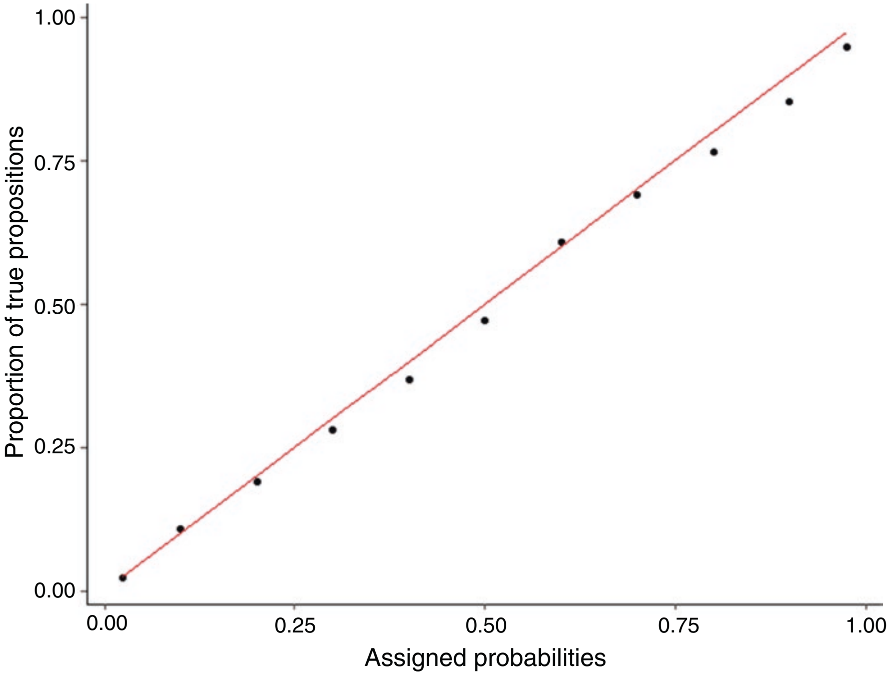

Fig. 3.1 Calibration graph for user 5265

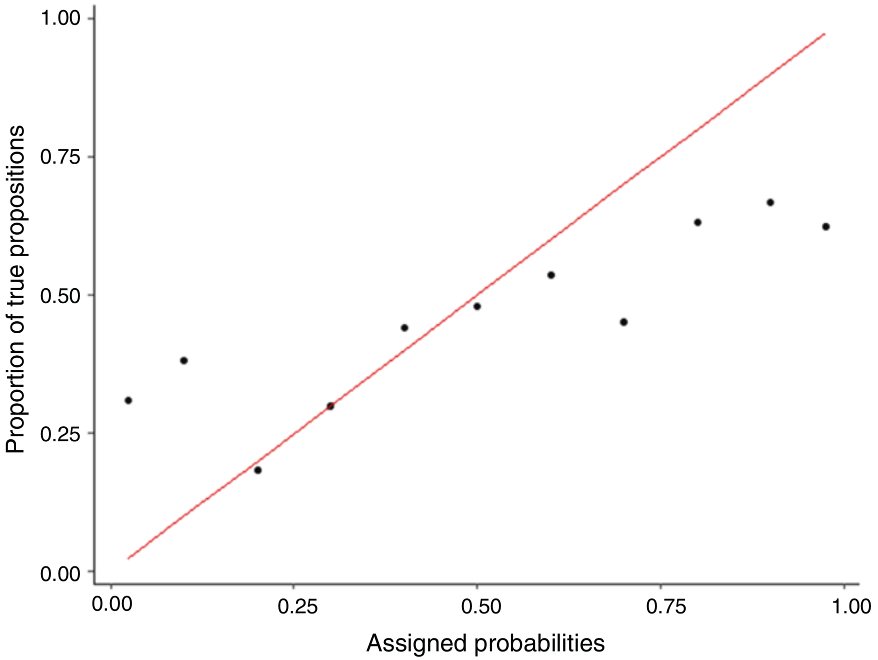

Fig. 3.2 Calibration graph for user 1890

What is worrying, however, is that we often simply do not know how calibrated humans are. We do not know whether, if someone is certain of something, that thing will actually be true 100% of the time or only 62% of the time like user 1890.

The force of the worry is even more apparent when we consider the importance of the decisions that rely on those judgments. On the basis of these judgments, a doctor may prescribe a risky medication, a jury may send a person to life in prison, or one country may invade another and send thousands, if not millions, to war or to an early death.

The situation is comparable to a pilot who flies blind in difficult weather conditions—making decisions on which their fate and the fate of others may depend—all while relying on judgments about what probably is out there, albeit without any knowledge about how accurate those judgments are.[1] Of course, few would feel comfortable with having a pilot who flies blind and relies on judgments about the world with unknown accuracy.

It is concerning, then, that society largely flies blind—that we do not know how accurate we are—in so many areas, including medicine, law, intelligence analysis, and others.

That said, there are some ways in which society does not fly blind, two of which I will mention here.

Chief among these is science—or perhaps I should say *high-quality* science. For example, there are many high-quality scientific studies that can reliably support the accuracy of judgments about, say, whether a particular medication is safe to take. Of course, science is not always an infallible guide to reality; poor-quality studies can be biased or misleading, for example. But the point is that, to a large extent, science helps us to avoid blind flight: it often helps to see the world for how it is so we can make better decisions about, for example, how to build rockets and send them to the moon, how to create vaccines that prevent the deaths of thousands, and how to predict and prepare for extreme weather events like storms.

Additionally, there are also a number of contexts where we do not fly blind because we have directly measured the accuracy of human judgment to determine how good it is. These contexts are the subject of the rest of this chapter, starting with cross-cultural psychology.

3.2 How Accurate Are Cultures: Inaccuracy in Cross-Cultural Psychology

Many studies have revealed that humans across cultures are generally inaccurate: people's confidence in something does not vary appropriately with the frequency with which that thing is true (e.g. see the review of Yates et al. (2002)).

[1] Technically, pilots typically have other instruments to navigate when they cannot visually see their environment, so the analogy is limited. But the point is that humans often do fly blind insofar as they are unaware of the accuracy of the judgments they rely on.

For example, one study examined the accuracy of two groups, each from different cultures (Lechuga & Wiebe, 2011).

Participants were asked to answer questions with various possible answers: for example, "Which animal runs faster?... (a) gazelle, (b) leopard." Participants selected an answer and then provided a probability estimate that their answer is correct. For example, they might have selected the answer "leopard" and assigned a 90% probability to it being correct.

The results found that participants were overconfident, although the extent to which this was the case varied between the groups. In one group from one culture, 100 students were quite overconfident when we consider specific categories of estimates. For instance, if we consider all of the propositions that anyone of them assigned a probability of 95% or more to, it turns out that only approximately 73% of them were true. In contrast, the other group of 108 students was more accurate but still somewhat overconfident. For example, if we consider all the propositions that they assigned a probability of 95% or more to, it turns out that approximately 87% of them were true. Of course, this study utilized a measure of unweighted binned calibration, so we do not know exactly what the distribution of inaccuracy for the individuals was (see Sect. 2.2.4.). The generalizability or external validity of the study is unclear too, but we can be confident that there was nevertheless some significant amount of inaccuracy among these two groups in any case.

Other studies have also reported the presence of overconfidence across various cultures (Lundeberg et al., 2000; Neyse et al., 2016; Whitcomb et al., 1995; Wright et al., 1978; Yates et al., 1996, 1998).

That said, these studies might not appear so concerning. After all, is it really that bad if someone is overconfident about whether a leopard is faster than a gazelle and other questions like this? One might think that even if humans are inaccurate about these trivial questions, they are more accurate in other more important contexts.

But alas, if only that were the case.

Instead, we have good evidence humans are inaccurate in important contexts too—including experts who give advice about real-world topics. Let us consider three examples of this: medicine, politics, and law.

3.3 How Accurate Are Medical Professionals: Inaccuracy in Medicine

Although it may sound dramatic, it is true that the importance of health cannot be overstated. The functioning and well-being of society depends on the health of its citizens. Plausibly, many people suffering severe illness can affirm its devastating impacts on multiple aspects of life: physical, social, financial, and emotional. And one of the hardest things a family can experience is the death of a loved one, especially if it is untimely and due to preventable health conditions.

Consequently, medical professionals have an important duty in maintaining the health of society. To do that, however, they need to have accurate judgments about a variety of matters: the causes of specific illnesses, the correct diagnoses of patients, appropriate treatments for illness, and how to correctly execute medical procedures—for example.

Given the importance of their work, one would hope that medical professionals would be highly accurate in their judgments.

However, various studies suggest this is not always the case.

Some of these suggest that medical students and medical professionals exhibit overconfidence in their knowledge.

One study, for example, examined the accuracy of judgments from 118 physicians in the United States (Meyer et al., 2013). They were asked to provide diagnoses that differed in their level of difficulty, as assessed by three experienced researchers. The easier cases were ones that most physicians correctly diagnosed; the harder cases were ones that most physicians incorrectly diagnosed. The concerning finding, however, was that they had similar levels of confidence about the correctness of their judgments in both kinds of cases. The physicians were asked to rate their confidence in the correctness of their diagnoses on a 11-point scale ranging from 0 to 10. The average confidence for both kinds of cases differed only slightly. They had an average confidence level of 7.2 regarding the easier cases for which 55.3% of the diagnoses were correct. Similarly, they had a confidence level of 6.4 regarding the more difficult cases for which merely 5.8% of the diagnoses were correct. We then have a 49.5% difference in accuracy accompanied by a mere 0.8 difference in confidence on an 11-point scale. On the basis of this and other findings, the study authors claimed that the physicians were generally "characterized by overconfidence in accuracy" (Meyer et al., 2013, p. 1952).

Another study examined the accuracy of physicians' judgments about cardiopulmonary variables (Perel et al., 2016). These judgments were important: they underpin important decisions which could affect their patients' lives. Despite that, the authors found that physicians had significantly inaccurate judgments about these variables. They claimed this demonstrates the "very limited clinical ability of physicians to correctly assess important physiological parameters" (Perel et al., 2016, p. 517).

Another kind of overconfidence has also been reported in medical students, one that refers to a discrepancy between one's estimated performance and their actual performance.

One study assessed the exam performance of 182 neuroanatomy students (Hall et al., 2016). The students also completed estimates of their own performance on these exams. They found a significant but relatively weak correlation between their actual performance and their self-perceived performance. The authors of the study conclude that it "provides further evidence to support the hypothesis that medical students are unable to accurately determine their own level of knowledge" (Hall et al., 2016, p. 493).

Another kind of overconfidence was reported in a study about how well 30 junior medical officers carried out particular medical procedures (Barnsley et al., 2004). More experienced medical practitioners assessed their performance and concluded that a significant number of them incorrectly carried out the tasks—either by omitting key steps or by wrongly executing others. Importantly, though, they did not find a significant correlation between the officers' confidence about how they did and how well they actually did according to the assessors. The authors conclude that this "raises serious concerns" about the officers' ability to "assess their own skills accurately" (Barnsley et al., 2004, p. 366).

Pessimistic findings like these led Westberg and Jason (1994) to conclude that "Medicine is dominated by unreflective doing" (p. 278), a sentiment that has been echoed decades later (Lam & Feller, 2020).

Of course, one problem with these studies is that they are often just single studies that focus on particular contexts that may be unrepresentative of the medical practice in general.

There are, however, more comprehensive reviews of the literature. For example, Berner and Graber (2008) conducted a review of the medical literature, and they concluded that "physicians in general underappreciate the likelihood that their diagnoses are wrong" (p. 52).

Other studies may also provide more insight into the accuracy of judgment, such as calibration studies for a variety of conditions, or studies about the prevalence of diagnostic errors in more natural settings.

Koehler et al. (2002) reviewed nine studies of the calibration of diagnoses and prognoses from physicians. They found that calibration varied from condition to condition; physicians were overconfident about some things and underconfident about others, but they were also sometimes relatively well calibrated. Importantly, however, they found that the biggest factor influencing calibration was the so-called *base rate*—that is, the frequency with which a given prognosis or diagnosis is true. For example, if a medical condition was very rare, then physicians were more likely to be *over*confident that it was the correct diagnosis, whereas if a condition was very common, then the physicians were more likely to be *under*confident that it was the correct diagnosis. Good calibration was achieved when the base rate coincided with the probabilities in the right way.

Koehler et al. (2002) used measures of calibration like those discussed in Chap. 2, and their results suggest physicians are often miscalibrated, at least for a range of conditions, even if they are well calibrated for others. If this is correct, then this suggests there are likely to be an excess of diagnostic errors in natural settings.

We can then triangulate these results with other studies that examine the prevalence of diagnostic error (even if they do not use the same calibration measures nor tell us anything about the calibration of medical judgment by themselves).

On this note, Arthur Elstein was a cognitive psychologist who studied medical decision-making for his entire career. He concluded that 10–15% of diagnoses are wrong (Elstein, 1995).

Since then, there have been a variety of research methodologies that investigate the prevalence of diagnostic error, each with their strengths and limitations.

One methodology involves autopsy studies which examine deceased patients and the possible causes of their deaths. In a review of the literature, Graber (2013) states that studies have identified major diagnostic errors in 10–20% of all cases. He cites as an example a study from Shojania et al. (2002) who reviewed 225 articles and concluded that 10.2% of the autopsy cases indicated a misdiagnosis which could have affected the outcome. Elsewhere, Leape et al. (2002) also state that autopsy estimates imply that, each year, 40,000 to 80,000 Americans die "preventable" deaths as a result of missed diagnoses—a death toll at least 13 times higher than the September 11th attacks. Interestingly, Shojania et al. (2002) also indicated a relative decrease of such diagnostic errors by 26.2% per decade, yet this was not statistically significant ($p = 0.1$); consequently, it is difficult to conclude whether this decrease represents a legitimate trend or merely a statistical fluke.

Regardless, one limitation with autopsy studies is that they do not detect misdiagnoses in those who have not died or have not been autopsied.

Physician reviews are then another way to study the prevalence of diagnostic error. In this methodology, physicians make diagnoses in various cases, and those cases are reviewed by other physicians to assess the probability of a diagnostic error.

In a study across 21 general practices, two reviewers agreed that missed diagnostic opportunities (MDOs) were present in 89 (4.3%) of 2057 consultations (Cheraghi-Sohi et al., 2021). Here, missed diagnostic opportunities were defined as "missed opportunities to make a correct or timely diagnosis based on the evidence available" (p. 977). That said, this study likely undercounts the number of misdiagnoses for two reasons: first, they only report data for 105 of the 207 cases where at least one physician identified an MDO, and second, each reviewer picked up on a large number of MDOs which the other did not, thereby suggesting that additional reviewers would have picked up on even more MDOs which the other two reviewers did not.[2] So while their study is useful, these two reasons suggest that their 4.3%

[2] According to the article, two physician reviewers examined 2,057 consultations, and there were "207 consultations identified by at least one reviewer as implicating an MDO" (p. 980). A number of these 207 consultations were subjected to joint review where both reviewers examined the same cases. For some reason, however, statistics are reported for only 105 of these 207 cases. The authors note that seven of them were omitted due to "administrative oversight" (p. 980–81), but this does not explain what happened to the other 95 cases. In any case, the two reviewers reportedly agreed that a missed diagnostic opportunity was present in 89 of the 105 reported cases, yielding their MDO rate of 4.3%, or 89 out of 2057.

This study's method then raises some concerns. For a start, consider the 207 cases where at least one reviewer thought a missed diagnostic opportunity was present. If statistics for only 105 of these are reported—only half!—and if the reviewers *both* agreed that missed diagnostic opportunities were present in 89 (86%) of those cases, then might they also have both agreed that there would have been even more missed opportunities in 86% of the 102 cases for which statistics are not reported. If this is the case, then they would have found around another 86 missed diagnostic opportunities, raising the total to around 175 missed opportunities, or 8.5% of the total consultations reviewed. If that is the case, then the missed diagnostic opportunity rate would have nearly doubled from the rate reported in their study. The concern, then, is that the 4.3% statistic is a substantial undercount of the missed opportunities the reviewers might have uncovered. That is the first concern.

The second concern, however, is that both reviewers picked up on many cases which the other reviewer did not. When the reviewers first examined the cases independently, they both agreed that there was a missed diagnostic opportunity in only 35 (16.9%) of the 207 cases. When they reviewed some of the other 207 cases together, however, they both agreed on an additional 54 cases (and perhaps even more, if statistics for the other 105 cases were reported). We can then ask the question: what would have happened if there was a *third* reviewer who reviewed the cases? Might they have identified another 50 or so cases of missed opportunities which the other reviewers would have agreed on if the third reviewer's cases were also subject to joint review? The worry is that there may have been—somewhat ironically—missed, missed diagnostic opportunities which would have been picked up on if there were more reviewers. For these reasons, the article's estimate of missed diagnostic opportunities appears to be an undercount, and possibly a drastic one.

missed diagnosis rate is an undercount. Instead, if the statistics for the other 102 cases were reported and if there had been more reviewers, the true diagnostic error rate may be even two or three times what was reported, somewhere between 8% and 15%—and maybe even more.

Another study reviewed three large observational studies and estimated that—similar to the previously discussed study—5.08% of US adults (a total of 12 million) will experience diagnostic errors annually (Singh et al., 2014). It is more difficult to assess the methodology of this paper, since details about, for example, inter-rater agreement are omitted from the article. In any case, their article may be susceptible to drastic underestimation for reasons similar to the above study.

Regardless, another meta-analysis appears to provide more optimistic estimates, at least at first glance. Gunderson et al. (2020) reviewed 22 studies of diagnostic errors. They concluded that at least 0.7% of people adults admitted to hospital experienced a harmful diagnostic error, with European and North American hospitals reporting lower rates than those in other continents.

However, their meta-analysis states that it differs from other estimates of diagnostic error (such as Singh et al.'s (2014) one above) for two reasons. First, it concerns a particular patient population: people who were admitted to hospital in contrast to those who are not. Second, it only examined studies of diagnostic error that used a particular methodology: more specifically, they examined only cases of "diagnostic errors" that involved patient harm—sometimes narrowly defined as harm requiring "prolonged length of stay or disability at the time of discharge" from the hospitals (p. 1014).

But of course, many diagnostic errors happen outside of hospital contexts and do not necessarily involve such adverse events, at least when they are so narrowly defined.

For that reason, this meta-analysis, while useful, is not exactly a generalizable measure of the prevalence of diagnostic error to other important medical contexts.

In any case, studies of diagnostic error using physician reviews are limited by another factor: the medical records that are reviewed by other physicians are often inaccurate. This has been demonstrated in studies involving "standardized patients"—that is, patients who are essentially undercover actors and present to medical clinics with real or fake conditions to assess the clinic's operations. Such studies have found that medical records are often inaccurate and that physicians often make misdiagnoses (Graber, 2013; Peabody et al., 2004).

We might then turn to other methodologies to investigate diagnostic error: patient and physician surveys. Graber (2013) states these kinds of studies consistently report that about half of respondents encountered diagnostic errors on at least a monthly basis. For instance, one study of 6,400 clinician respondents indicated that nearly half reported seeing errors monthly (MacDonald, 2011). Another study found that more than half of 726 pediatricians in the United States reported making a diagnostic error at least once a month (Singh et al., 2010), with patient harm being "not uncommon" (p. 70).

Another survey used random digit dialing to contact 1,207 members of the US public, 42% of whom reported "that they experienced an error in their own care or that of a family member," with 24% reporting errors with "serious" consequences and 10% reporting errors which resulted in death (Blendon et al., 2002, p. 1934–35). Additionally, 821 physicians also completed the same survey and mailed in their responses; these responses were slightly more optimistic but still worrying, with 35% indicating they or a family member experienced a medical error and 7% reporting errors that resulted in death. That said, the study focused on "medical" errors which were preventable and not necessarily just on "diagnostic" errors that may or may not have been preventable.

Another study—not discussed by Graber—surveyed 2,201 respondents and found that 35% of them reported experiencing a misdiagnosis, either directly or through a friend or family member (TriMed Staff, 2006). Respondents further said that, of those errors, 25% resulted in permanent harm or death. Commenting on the survey, Professor of medicine Robert Watcher provides a sobering assessment:

> Today in America, hundreds of patients will be falsely reassured and panicked, and many of them will be medicated, scanned, and even cut open because of the wrong diagnosis. (TriMed Staff, 2006)

Studies like these lead the *National Academies of Sciences, Engineering, and Medicine* (2015) to write, "[i]t is likely that most people will experience at least one diagnostic error in their lifetime, sometimes with devastating consequences" (p. 355).

Graber (2013) shares similarly pessimistic comments regarding the studies of diagnostic error:

> *In summary*, a wide range of different research approaches have been used to estimate diagnostic error rates, all suggesting that the incidence is unacceptably high. Although true incidence data are lacking, a wide variety of research studies suggest breakdowns in the diagnostic process result in a staggering toll of harm and patient deaths (p. ii25)

What these studies suggest is that although physicians and medical professional have accurate judgments a lot of time—and perhaps even in a large majority—there is a substantial remainder where inaccuracy and overconfidence are present.

Furthermore, such inaccuracy is not a trivial matter: it has important implications for the lives of arguably millions of people. And plausibly, for some ill patients at least, judgmental accuracy is literally a matter of life and death.

3.4 How Accurate Are Political Experts: Inaccuracy in Political Judgment

Overconfidence has also been found in judgments about political matters—such as the outcomes of wars, elections, trade agreements, and other international affairs. In a groundbreaking 20-year study, hundreds of people made forecasts (i.e., predictions) about future political events (Tetlock, 2005). There, 284 of the people making these forecasts were experts. In this context, an expert was defined as "a professional who makes his or her livelihood by commenting or offering advice on political and economic trends of significance to the well-being of particular states, regional clusters of states, or the international system as a whole" (Tetlock, 2005, p. 239).

As with medical experts, we would hope that the expertise of these political experts ensures that they can make accurate judgments. And this is especially so since their opinions occasionally inform the important opinions and decisions, such as those from governmental and nongovernmental organizations.

So, then, how well did these experts do?

Again, poorly—at least when considering the long-term predictions of one group of experts. Out of all the propositions they assigned an 80% probability to, only 45% of them were true, and the other 65% were false. And out of all the propositions they were virtually certain about—the propositions they assigned a probability of 1 or thereabouts—approximately 30% of them were *false*. Additionally, 19% of the things they were *completely certain* would *not* happen actually *did* happen.

Of course, those who did so poorly were not *all* of the experts in the study. It was just an inaccurate group of them. So far as I can tell, the study author, Philip Tetlock, did not state how big this group was, but another source indicates it was possibly 189 experts in total (Rieber, 2004).

In any case, we might ask what distinguishes the more accurate experts from the less accurate experts. Was it their education, their line of work, their years of experience, or how relevant their expertise was to the question at hand? After all, one might think that these correlate with better accuracy.

Perhaps surprisingly, none of these particular metrics correlated substantially with accuracy. As Tetlock (2005) notes:

> It made virtually no difference whether participants had doctorates, whether they were economists, political scientists, journalists, or historians, whether they had policy experience or access to classified information, or whether they had logged many or few years of experience in their chosen line of work. (p. 68)

There were a couple of exceptional variables though.

One variable correlated particularly well with *inaccuracy*: fame. As Tetlock (2005) notes, "better known forecasters—those more likely to be fêted by the media—were less well calibrated than their lower-profile colleagues" (p. 68). The idea behind this is as follows. People who are less accurate and more overconfident are more likely to jump to conclusions and have extreme opinions. The media then often seeks out these people to provide interesting, informative, or counterbalancing opinions on subjects that the public cares about. Consequently, they become more famous. The result is that those with greater fame have more overconfidence, more extreme opinions, and less accuracy—or so that was what Tetlock reported.

But another variable correlated *positively* with accuracy: particular thinking styles. There were some less well-known experts who did quite well, at least when it comes to predictions about the short-term future. Tetlock says that they exhibited the thinking style of "foxes." That is, they were:

> thinkers who know many small things (tricks of their trade), are skeptical of grand schemes, see explanation and prediction not as deductive exercises but rather as exercises in flexible

> "ad hocery" that require stitching together diverse sources of information, and are rather diffident about their own forecasting prowess, and… rather dubious that the cloudlike subject of politics can be the object of a clocklike science (pp. 74–75)

In contrast, those who did poorly—the ones alluded to above—are what Tetlock called "hedgehogs." He states they were:

> thinkers who "know one big thing," aggressively extend the explanatory reach of that one big thing into new domains, display bristly impatience with those who "do not get it," and express considerable confidence that they are already pretty proficient forecasters, at least in the long term (p. 73)

The idea is that such inaccurate experts have thinking styles that make them jump to conclusions and blind them to their own inaccuracy. Their status as an "expert" then gives them a false sense of confidence in their own mistaken views, making them less reliable in their opinions than other experts in the area. The result of this is the so-called *illusion of knowledge*: the illusion that one's knowledge entitles them to be confident about more things than what they actually know. As Tetlock puts it, then, "it should be downright disturbing to discover that knowledge handicaps so large a fraction of forecasters" (p. 81).

So we have once again seen evidence of inaccuracy in another important context—this time, political judgments. We have also discussed Tetlock's views about what correlates with accuracy, although we will return to consider correlates of forecasting accuracy in Chap. 7. There, we will see that the evidence does indeed show thinking styles correlate with accuracy (albeit not exactly along the "Fox/Hedgehog" lines that Tetlock (2005) discussed).

For now, though, we will turn to consider accuracy in another important context: law.

3.5 How Accurate Are Judges and Juries: Inaccuracy in Law

Ideally, the law plays an important role in maintaining a just and flourishing society, one that punishes crime, thereby reducing it and promoting justice.

Yet human judgment also plays an important role in the law. This is true at various levels. It is true when we consider the politicians who use their judgment to estimate which laws are best for society and who then make law accordingly. But it is also true in the courts where the law is administered and where, hopefully, justice is served.

Much attention has been devoted to the accuracy of judgments from judges or juries who pronounce people's criminal guilt—or lack thereof—in various cases. When a judge or jury makes a judgment of guilt that is rejected in subsequent judicial investigations, this is called an *exoneration*.

Exonerations are often indicators of inaccurate judgments or false convictions, although many false convictions are probably never discovered, so they do not result in exonerations.

What follows, then, are some potentially interesting statistics about exonerations. Exonerations are fairly common; as of 2015, there were about three exonerations per week in the United States, most involving rape or murder (Gross, 2017). There were 1,900 exonerations between 1989 and 2016: out of those, 76% were exonerations of convictions from juries, and 7% were convictions from judges; 23% were cleared with DNA evidence, while 77% were not; and as a conservative estimate, 38% of exonerated people spent 10–39 years in prison, with the average person spending 9 years in prison for a crime where they were ultimately found to be legally innocent (Gross, 2017).

The rate of exoneration differs from crime to crime: murder and rape cases are more likely to be exonerated than other crimes—murder cases because their cases receive greater attention, resources, and reinvestigation and rape cases because they are more likely to have vindicating DNA evidence compared to many other crimes.

The factors that contribute to false convictions in exoneration cases also vary from crime to crime. For rape and robbery, mistaken eyewitness identifications are the most commonly cited contributing factor; for child sex abuse cases, it is perjury or false accusations; for drug crimes, it is misleading forensic evidence, such as substances that are misclassified as illegal drugs; and for murders, it is both official misconduct and perjury. Gross (2017) thinks racism also plays a role in some false convictions, especially since "[o]ther things equal, an African American is about six times more likely to be falsely convicted of a crime and then exonerated than a non-Hispanic white American" (p. 778).

So ultimately, the causes of false convictions can vary from context to context. This lead Samuel Gross to say that "[i]t makes no more sense to talk about the 'leading cause' or even the 'causes' of 'false conviction' in general than it does to talk about the causes of 'diseases'" (Gross, 2017, p. 772).

Having explored the causes of false convictions, then, we might ask the question: what is the false conviction rate? How accurate are the judgments of guilt from judges or juries?

Unfortunately, there is currently no way to know the answer to this question with a high degree of precision and confidence. A part of the reason for this is that the vast majority of convictions are never seriously re-investigated in order for false convictions to be identified.

That said, some types of crimes are much more likely to be reinvestigated than others, with one type being by far the most likely: crimes resulting in death sentences.

The rate of exonerations for death sentences in the United States is much higher than for any other kind of criminal conviction. This is largely because death sentences receive more attention, resources, and reinvestigation before and after the conviction. This also explains why defendants who are on death row but then receive life sentences are much less likely to be reinvestigated and reconsidered, thus drastically lowering their chances of exoneration.

Nearly all death sentences come from trials by juries, so they potentially provide a better estimate of the accuracy of juries than judges.

To study their accuracy, Gross et al. (2014) examined death-sentence exonerations in the United States and estimated the false conviction rate for death sentences. They studied 7,482 defendants who were sentenced to death from 1973 to 2004, 1.6% of which (117) were exonerated, while the others either were executed

(12.6%), died on death row from other causes like suicide (4%), were still on death row (46.1%), or were removed from death row but not exonerated (35.8%). Using this data, they focused on the proportion of cases that received a high degree of attention, and they estimated that 4.1% of death sentence convictions from 1973 to 2004 were false, meaning that some of the 1,320 defendants executed since 1973 were innocent. That said, they also think this is a "conservative estimate" which is likely an "undercount" (Gross et al., 2014, p. 7243).

The 4.1% estimate is also similar to that of Risinger's (2006) investigation: he estimated a false conviction rate of 3.3% to 5% for death sentences in the 1980s (although he focus only on those cases where the victim was raped and then murdered).

Gross et al.'s (2014) article makes some unsettling claims given that their estimate of 4.1% differs from 1.6% of people who were actually exonerated. In particular, they claim their study provides the "disturbing news that most innocent defendants who have been sentenced to death have not been exonerated, and many—including the great majority of those who have been resentenced to life in prison—probably never will be" (p. 7235).

They think, however, that further disturbing news comes from the fact that defendants are more likely to be sentenced to life imprisonment if juries have doubts about their guilt. They claim that if the juries are more likely to doubt a defendant's guilt, then the defendant is more likely to be innocent. But if that is the case, and if they are more likely to be innocent, then they are also more likely to be sentenced to life in prison rather than death. So, if that is the case, then they are less likely to reinvestigated and less likely to be exonerated. The upshot of all this is that if Gross et al. (2014) are correct, then convicted defendants who are more likely to be innocent are less likely to have their innocence proven and are less likely to be exonerated. As they say:

> The net result is that the great majority of innocent defendants who are convicted of capital murder in the United States are neither executed nor exonerated. They are sentenced, or resentenced to prison for life, and then forgotten" (p. 7235)

So, some scholars think the study of false convictions in death sentences provides saddening news. If a reliable conservative estimate of the false conviction rate is 4.1%, then this is obviously too high. Of course, as discussed in Chap. 2, perfect judgmental accuracy permits some degree of error: even if a perfectly accurate person was 99% confident that a defendant committed a crime, the defendant would be innocent 1% of the time. But if the false conviction rate is 4% or 5%, then juries would be perfectly accurate only if they were about 95% confident that the defendant committed such serious crimes, meaning the criminal would be innocent 1 out of 20 times. Of course, it is debatable whether any threshold of confidence warrants a death sentence, but it should be clear that, at the very least, a confidence level of around 95% does not. So there is likely overconfidence among jury members, at least when we consider the false conviction rate for death sentences.

What, then, does this tell us about the false conviction rate in general, including for other crimes?

Gross et al. (2014) say it is difficult to extrapolate from death sentence crimes to others. This is because authorities are more likely to pursue defendants in court with weak evidence if people have been killed, potentially leading to a higher proportion of false convictions. Despite that, Gross (2017) does say that "with a 4% error rate for death sentences, it's hard to believe that false convictions occur in a mere fraction of a percent of lesser cases" (p. 769). He also thinks the 4.1% estimate "suggests that the rate for other violent felonies is somewhere in the range from one to several percent" (p. 785).

Not all agree with Gross, however. Lawyer Paul G. Cassell (2018) argues that the false conviction rate is far, far lower—as low as 0.016–0.062%. This is a surprising calculation for two reasons. First, the exoneration rate for Gross et al.'s (2014) sample of death sentences alone was much higher at 1.6% (bearing in mind that this rate is lower than the 4.1% estimate since many death sentences were converted to life sentences, thus reducing the likelihood of their reinvestigation and exoneration). Second, if Cassell is right, and if the jurors are perfectly calibrated, then they would need to have an average confidence level of at least 99.94% per conviction. Calibration like this seems unlikely, especially since other evidence suggests miscalibration is very pervasive. For these two reasons, Cassell's claims are suspicious.

However, the false conviction rate is also a politicized topic. Zalman and Norris (2021) claim that "all who write about innocence in the criminal legal system (or about other legal or social topics), including ourselves, are influenced by ideological preferences" (p. 614). For example, they claim Gross is an avowed liberal who argued against capital punishment in the 1990s, and his later work provided important data for the "Innocence Project," a nonprofit organization to free the innocent and prevent false conviction. They also claim Cassell is a conservative lawyer who has a history of making questionable arguments for specific policies. In particular, they cite his arguments that there is a very low rate of false convictions that are based on false confessions (Cassell, 1997). They also refer to Leo and Ofshe's (1997) critique of those arguments. Leo and Ofshe made a number of methodological criticisms of Cassell (1997), including that key quantities to estimate his rate are unknown (such as the number of confessions) and that numerous false confessions are unreported—among other topics. Leo and Ofshe then claim "Cassell's quantification scheme is merely a rhetorical device to permit him to argue for the superiority of his policy preferences" (p. 563). They further seem to suggest "we must necessarily reject his conclusions as ideologically driven and the products of a commitment to the advocacy of a particular value position rather than to the empirical ascertainment of truth" (p. 562–563). Zalman and Norris (2021) say that they "find Leo and Ofshe's critique persuasive" (p. 614).

For these reasons, it is not clear how many scholars take Cassell's arguments seriously. Nevertheless, it is suspicious that Cassell's estimate is drastically lower than even just the bare exoneration rate for death sentences which itself is likely an undercount of the true prevalence of false convictions for reasons discussed by Gross et al. (2014).

3.6 Other Evidence of Inaccuracy: Disagreement

Of course, the aforementioned studies are arguably not the only evidence of inaccuracy. Another source of evidence is disagreement. It is obvious that when two people disagree on a question, each espousing incompatible answers, at least one of them must be wrong.

And yet disagreement proliferates throughout society. One would hope that this disagreement is quickly and efficiently resolved. Perhaps science is like this—at least some of the time. But in other contexts, it is clear that it is not like this. We can see this when we consider topics in religion or politics, for example. There, millions of people disagree with each other about, for example, whether there is a God who was embodied in Jesus, whether gun control saves lives or whether nuclear power is safe. Furthermore, these issues are potentially important: our judgments about religious or political issues can determine whether one aborts a baby or brings a new life into existence, or whether guns are banned, thereby potentially reducing mass shootings and saving lives. So millions of people disagree with each other, each believing opposing things which cannot be simultaneously true. And even if some of these groups of people are right, the remainder must be wrong, thereby indicating that inaccuracy, at least on some measures of it, is both pervasive and tremendously consequential (Kahneman et al., 2021). This, in fact, is closely related to Kahneman et al.'s book *Noise* (2021): there, they argue that there is an unacceptably high amount of "noise" or disagreement—and hence inaccuracy—in medicine, in forecasts, in child custody decisions, in forensic science, and in many other areas—basically "wherever there is judgment".

3.7 Contexts with Underconfidence

But this chapter also makes it clear that not everybody is always equally overconfident—take Tetlock's fox-like forecasters, for example.

There are also various contexts in which people can be underconfident in their judgments. One concerns easy questions. Occasionally, people are underconfident in easy questions but overconfident in hard questions. This is called the *hard-easy* or *difficulty effect* (Yates et al., 2002).

Another context concerns the forecasts of particular intelligence analysts. Mandel and Barnes (2014) studied the accuracy of 1,514 intelligence forecasts from various Canadian intelligence reports. They found that the forecasts were very well calibrated, although there was some degree of inaccuracy. In particular, the forecasts were often *underconfident*, especially in forecasts about harder questions or more important questions. For example, the forecasts assigned 70% probabilities to events that happened closer to 80% of the time (Mandel & Barnes, 2014). So that is one context where underconfidence may be present.

It is also possible that sex and gender play a role in other contexts. One study found that, sadly, female students were more likely to underestimate their performance on neuroanatomy examinations (Hall et al., 2016). Another study reported that female students underestimated their mid-clerkship performance, even though they outperformed their male counterparts (Lind et al., 2002). In saying that, two other studies did not find similar results. One study failed to find a statistically significant difference among gender assessments of self-competence (Minter et al., 2005). Another study found that women were more likely to underestimate their surgical skills, but this difference disappeared when other variables were taken into account (De Blacam et al., 2012). The authors claimed that "although women did underestimate their competency, the trend cannot be ascribed to gender" (p. 730). Consequently, the studies on gender show mixed findings about underconfidence.

3.8 Summary

So we have reviewed evidence from different fields, including studies of general knowledge across cultures, of medical judgment, of political judgment, and of criminal convictions in law.

In general, the picture they paint is one of inaccuracy. Across these fields, people are often overconfident in the correctness of their judgments.

But there are also some contexts where individuals are underconfident. Some studies have found that people can even be underconfident about especially easy questions. Also, some intelligence analysts are well calibrated, albeit somewhat underconfident. There is also mixed evidence that gender may play a role too, with some studies indicating that women are underconfident in their knowledge and abilities while other studies paint a different picture.

But it is misleading to say that humans are inaccurate simpliciter.

This is for two reasons.

First, there are many contexts where we have judgments that are perfectly accurate, at least for our purposes. For example, people typically have accurate judgments about where they live, whether they need to breathe oxygen, and a myriad other things where our accuracy is so commonplace and boring that we simply fail to even notice it. Meteorologists are also often pretty accurate at predicting the weather, potentially because they rely heavily on statistics, although some studies also suggest they are still somewhat miscalibrated for more rare events like extreme storms (see a review in Koehler et al. (2002)). Regardless, we have not focused as much on these contexts of accuracy. This is for several reasons: there are unlikely to be studies announcing mundane facts that, say, humans are perfectly accurate and calibrated when recalling where they live; readers of this book are unlikely to be interested in these mundane facts; and, most importantly, I think society is likely to benefit more from focusing on areas for improvement—on areas where we are inaccurate in ways that can be mitigated.

Regardless, the second reason it is misleading to say that humans are simply inaccurate is that even in the contexts where humans are worryingly inaccurate, they still get it right to some extent. The available evidence suggests that when doctors, juries, or humans are 100% certain of things, they are often still right most of the time, whether that is 60% of the time, 90% of the time or somewhere in between.

But it is genuinely worrying, however, that humans are not perfectly accurate: that, say, doctors may be certain that some patients have diseases that they do not have, that political experts will be certain of some events that will not happen, and that juries may be certain of the guilt of some innocent people. The result of this inaccuracy is significant: thousands of people will fall sick and die of preventable conditions, thousands may die in wars waged on false premises, and hundreds of innocent people will be executed or sent to life in prison.

Yet humans are not simply accurate or inaccurate. Instead, their accuracy varies from context to context; this is the key idea of the context-dependent model of accuracy. In some contexts, humans are perfectly accurate; in other contexts, they are disturbingly inaccurate. But in many more contexts, we simply do not know, merely because accuracy is not reliably assessed.

However, the inaccurate contexts provide little ground for unquestioning optimism, and they make it clear that we can truthfully add various qualifiers to our claims about inaccuracy: for example, humans are often much less accurate than we would hope or expect, with emphasis on qualifying words like "often," "much less," and so forth.

We might then ask what explains why we are so inaccurate. This question is spoken to by the following three chapters of this book: one on so-called *metacognition*, another on the so-called *heuristics* by which people often arrive at their judgments, and another chapter on the *evolution* of our psychology. In the fourth chapter on metacognition, we will see that people often have inaccurate judgments about their own accuracy; this can partially explain why inaccurate judgments persist. In the fifth chapter on heuristics, we will see that people often employ heuristics which produce biased judgements; this can partially explain in more detail how such inaccurate judgments develop. In the sixth chapter on evolution, we will examine accounts of how our fallible psychology is somewhat expectable given our origins. If true, these evolutionary accounts would partially explain why our minds and reasoning processes are so sub-optimal for truth-seeking in the first place.

References

Barnsley, L., Lyon, P. M., Ralston, S. J., Hibbert, E. J., Cunningham, I., Gordon, F. C., & Field, M. J. (2004). Clinical skills in junior medical officers: A comparison of self-reported confidence and observed competence. *Medical Education, 38*(4), 358–367. https://doi.org/10.1046/j.1365-2923.2004.01773.x

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *American Journal of Medicine, 121*(5 SUPPL), S2–S23. https://doi.org/10.1016/j.amjmed.2008.01.001

Blendon, R. J., Desroches, C. M., Rosen, A. B., & Herrmann, M. J. (2002). Views of practicing physicians and the public on medical errors. *The New England Journal of Medicine, 8*, 1933.

Cassell, P. G. (1997). Protecting the innocent from false confessions and lost confessions–and from miranda. *Journal of Criminal Law and Criminology, 88*, 497.

Cassell, P. G. (2018). Overstating America's wrongful conviction rate: Reassessing the conventional wisdom about the prevalence of wrongful convictions. *Arizona Law Review, 60*(4), 815–864.

Cheraghi-Sohi, S., Holland, F., Singh, H., Danczak, A., Esmail, A., Morris, R. L., Small, N., Williams, R., de Wet, C., Campbell, S. M., & Reeves, D. (2021). Incidence, origins and avoidable harm of missed opportunities in diagnosis: Longitudinal patient record review in 21 English general practices. *BMJ Quality & Safety, 30*(12), 977–985. https://doi.org/10.1136/bmjqs-2020-012594

De Blacam, C., O'Keeffe, D. A., Nugent, E., Doherty, E., & Traynor, O. (2012). Are residents accurate in their assessments of their own surgical skills? *American Journal of Surgery, 204*(5), 724–731. https://doi.org/10.1016/j.amjsurg.2012.03.003

Elstein, A. (1995). Clinical reasoning in medicine. In J. Higgs (Ed.), *Clinical reasoning in the health professions* (pp. 49–59). Butterworth-Heinemann Ltd..

Graber, M. L. (2013). The incidence of diagnostic error in medicine. *BMJ Quality & Safety, 22*(Suppl 2), ii21–ii27. https://doi.org/10.1136/bmjqs-2012-001615

Gross, S. R. (2017). What we think, what we know and what we think we know about false convictions. *Ohio State Journal of Criminal Law, 14*(2), 753–786.

Gross, S. R., O'Brien, B., Hu, C., & Kennedy, E. H. (2014). Rate of false conviction of criminal defendants who are sentenced to death. *Proceedings of the National Academy of Sciences, 111*(20), 7230–7235. https://doi.org/10.1073/pnas.1306417111

Gunderson, C. G., Bilan, V. P., Holleck, J. L., Nickerson, P., Cherry, B. M., Chui, P., Bastian, L. A., Grimshaw, A. A., & Rodwin, B. A. (2020). Prevalence of harmful diagnostic errors in hospitalised adults: A systematic review and meta-analysis. *BMJ Quality & Safety, 29*(12), 1008–1018. https://doi.org/10.1136/bmjqs-2019-010822

Hall, S. R., Stephens, J. R., Seaby, E. G., Andrade, M. G., Lowry, A. F., Parton, W. J. C., Smith, C. F., & Border, S. (2016). Can medical students accurately predict their learning? A study comparing perceived and actual performance in neuroanatomy. *Anatomical Sciences Education, 9*(5), 488–495. https://doi.org/10.1002/ase.1601

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment. Little, Brown Spark.

Koehler, D. J., Brenner, L., & Griffin, D. (2002). *The calibration of expert judgment: Heuristics and biases beyond the laboratory. Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 686–715).

Lam, J., & Feller, E. (2020). Are we right when We're certain? Overconfidence in medicine. *Rhode Island Medical Journal, 103*(2), 11–12.

Leape, L. L., Berwick, D. M., & Bates, D. W. (2002). Counting deaths due to medical errors—Reply. *JAMA, 288*(19), 2405. https://doi.org/10.1001/jama.288.19.2405-JLT1120-2-3

Lechuga, J., & Wiebe, J. S. (2011). Culture and probability judgment accuracy: The influence of holistic reasoning. *Journal of Cross-Cultural Psychology, 42*(6), 1054–1065. https://doi.org/10.1177/0022022111407914

Leo, R. A., & Ofshe, R. J. (1997). Using the innocent to scapegoat Miranda: Another reply to Paul Cassell. *Journal of Criminal Law & Criminology, 88*, 557.

Lind, D. S., Rekkas, S., Bui, V., Lam, T., Beierle, E., & Copeland, E. M. (2002). Competency-based student self-assessment on a surgery rotation. *Journal of Surgical Research, 105*(1), 31–34. https://doi.org/10.1006/jsre.2002.6442

Lundeberg, M. A., Fox, P. W., Brown, A. C., & Elbedour, S. (2000). Cultural influences on confidence: Country and gender. *Journal of Educational Psychology, 92*(1), 152–159. https://doi.org/10.1037/0022-0663.92.1.152

MacDonald, O. W. (2011). *Physician perspectives on preventing diagnostic error*.

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America, 111*(30), 10984–10989. https://doi.org/10.1073/pnas.1406138111

Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Internal Medicine, 173*(21), 1952–1961. https://doi.org/10.1001/jamainternmed.2013.10081

Minter, R. M., Gruppen, L. D., Napolitano, K. S., & Gauger, P. G. (2005). Gender differences in the self-assessment of surgical residents. *American Journal of Surgery, 189*(6), 647–650. https://doi.org/10.1016/j.amjsurg.2004.11.035

National Academies of Sciences, E. (2015). *Improving Diagnosis in Health Care*. https://doi.org/10.17226/21794

Neyse, L., Bosworth, S., Ring, P., & Schmidt, U. (2016). Overconfidence, incentives and digit ratio. *Scientific Reports, 6*(1), 1–8. https://doi.org/10.1038/srep23294

Peabody, J. W., Luck, J., Jain, S., Bertenthal, D., & Glassman, P. (2004). Assessing the accuracy of administrative data in health information systems. *Medical Care, 42*(11), 1066–1072. https://doi.org/10.1097/00005650-200411000-00005

Perel, A., Saugel, B., Teboul, J. L., Malbrain, M. L. N. G., Belda, F. J., Fernández-Mondéjar, E., Kirov, M., Wendon, J., Lussmann, R., & Maggiorini, M. (2016). The effects of advanced monitoring on hemodynamic management in critically ill patients: A pre and post questionnaire study. *Journal of Clinical Monitoring and Computing, 30*(5), 511–518. https://doi.org/10.1007/s10877-015-9811-7

Rieber, S. (2004). Intelligence analysis and judgmental calibration. *International Journal of Intelligence and CounterIntelligence, 17*(1), 97–112. https://doi.org/10.1080/08850600490273431

Risinger, D. M. (2006). Innocents convicted: An empirical justified factual wrongful conviction rate. *Journal of Criminal Law & Criminology, 97*, 761.

Shojania, K. G., Burton, E. C., McDonald, K. M., & Goldman, L. (2002). The autopsy as an outcome and performance measure. *Evidence Report/Technology Assessment (Summary), 58*, 1–5.

Singh, H., Thomas, E. J., Wilson, L., Kelly, P. A., Pietz, K., Elkeeb, D., & Singhal, G. (2010). Errors of diagnosis in pediatric practice: A multisite survey. *Pediatrics, 126*(1), 70–79. https://doi.org/10.1542/peds.2009-3218

Singh, H., Meyer, A. N. D., & Thomas, E. J. (2014). The frequency of diagnostic errors in outpatient care: Estimations from three large observational studies involving US adult populations. *BMJ Quality & Safety, 23*(9), 727–731. https://doi.org/10.1136/bmjqs-2013-002627

Tetlock, P. (2005). *Expert political judgment: How good is it? How can we know?* Princeton University Press.

TriMed Staff. (2006). *Study: 1 in 6 Americans has suffered from a medical misdiagnosis*. Health IT. https://www.healthimaging.com/topics/health-it/study-1-6-americans-has-suffered-medical-misdiagnosis

Westberg, J., & Jason, H. (1994). Fostering learners' reflection and self-assessment. *Family Medicine, 26*(5), 278–282.

Whitcomb, K. M., Önkal, D., Curley, S. P., & George Benson, P. (1995). Probability judgment accuracy for general knowledge. Cross-national differences and assessment methods. *Journal of Behavioral Decision Making, 8*(1), 51–67. https://doi.org/10.1002/bdm.3960080105

Wright, G. N., Phillips, L. D., Whalley, P. C., Choo, G. T., Ng, K.-O., Tan, I., & Wisudha, A. (1978). Cultural differences in probabilistic thinking. *Journal of Cross-Cultural Psychology, 9*(3), 285–299. https://doi.org/10.1177/002202217893002

Yates, J. F., Lee, J. W., & Shinotsuka, H. (1996). Beliefs about overconfidence, including its cross-national variation. *Organizational Behavior and Human Decision Processes, 65*(2), 138–147. https://doi.org/10.1006/obhd.1996.0012

Yates, J. F., Lee, J. W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence? *Organizational Behavior and Human Decision Processes, 74*(2), 89–117. https://doi.org/10.1006/obhd.1998.2771

Yates, J. F., Lee, J.-W., Sieck, W. R., Choi, I., & Price, P. C. (2002). Probability judgment across cultures. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 271–291). Cambridge University Press.

Zalman, M., & Norris, R. J. (2021). Measuring innocence: How to think about the rate of wrongful conviction. *New Criminal Law Review, 24*(4), 601–654.

Chapter 4
How We Evaluate Our Thinking: The Accuracy of Our Metacognition

The preceding chapter indicated that people are often inaccurate in their judgments. Of course, such judgments involve human *cognition*—that is, the thoughts that people have.

Beyond that, though, there is a sense in which people are frequently also inaccurate in their *metacognition*—that is, in the thoughts that they have about their own thoughts. More specifically, they are often inaccurate when they estimate their own accuracy, especially in relation to other people.

This can in turn explain why inaccuracy persists: after all, if our judgments are often inaccurate, it would not be surprising to see this inaccuracy persist if we also have inaccurate judgments about just how inaccurate we really are.

We will then review evidence of this so-called *metacognitive inaccuracy* in this chapter, followed by various potential explanations of why it happens.

4.1 Evidence of Metacognitive Inaccuracy

To some extent, metacognitive inaccuracy was obvious from the studies surveyed in the previous chapter. In some of them, such as Lechuga and Wiebe's (2011) study, participants had to give judgments in the form of their best guesses about the answers to questions. Then, they had to give their probability estimates that those answers were correct. And there is a sense in which we might think of these estimates as metajudgments—that is, as judgments about judgments. After all, they made a judgment in the form of a guess about which answer is correct, and then they made a probabilistic judgment about that judgment being correct. And if that is the case, then one might conclude that such metajudgments were inaccurate. To that extent, we might interpret much of the literature on accuracy as evidence for the inaccuracy of our metacognition.

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0_4

But even if we do not interpret it this way, we did not need to look at this literature to find evidence of inaccurate metacognition.

Instead, there is already a large literature on the inaccuracy of metacognition. The studies in this literature often focus on what is called the *unskilled-and-unaware effect*, also known as the *Dunning-Kruger effect*. Put simply, this refers to when one is both unskilled in a particular domain and unaware of just how unskilled they are. In our context, though, perhaps the term "inaccurate" is more apt than "unskilled," since people are obviously inaccurate in many contexts, but it is not obvious that being inaccurate is straightforwardly the same as lacking a skill.

Regardless, academic performance is often cited in support of the effect. Numerous studies have found that underperforming students overestimate how well they do on academic tests, thereby lacking self-awareness of their own underperformance (Callender et al., 2016; Ehrlinger et al., 2008; Hall et al., 2016).

However, there have been challenges to the existence of this effect.

In particular, Miller and Geraci (2011) suggested that the unskilled are not so unaware. They base their suggestion on the finding that some underperforming groups also reported less confidence in their judgments about how well they did compared to those who overperformed.

In saying that, another study from Händel and Dresel (2018) critiqued this suggestion. Like Miller and Geraci (2011), they found that underperforming students were less confident in their assessments of their own accuracy. However, they found that this confidence about their own judgments did not correlate with the accuracy of those judgments. In other words, they were likely to be similarly as confident about their wrong answers as compared to their right ones. For this reason, Händel and Dresel (2018) concluded that underperforming students had lower levels of confidence for irrational reasons. One such reason is having a lower level of confidence by default, one that is fairly insensitive to their own actual performance. In any case, Händel and Dresel failed to find support for the notion that the underperforming students possessed some previously unappreciated awareness of when they were right versus when they were wrong.

Nevertheless, the literature generally finds that underperforming students tend to overestimate their competence.

But the studies on the Dunning-Kruger effect also reveal a paradox that is in tension with this finding. Kruger and Dunning (1999) asked participants in their studies to rank how well they did *relative to others*. When they did this, they found that, perhaps unsurprisingly, underperforming participants ranked themselves higher than they really were. But what is more surprising is that higher performing participants also ranked themselves *lower* than they really were.

At first, this might seem like evidence of a kind of *under*confidence.

However, such higher performing participants were still overconfident in their over competence, a finding replicated in a different study by Händel and Dresel (2018). Instead, Kruger and Dunning suggested that such participants merely underestimate how competent they are *relative* to others—not in absolute terms. They initially suggested that this was not an expression of underconfidence, but rather of the *false consensus effect*: the tendency to think that others are more similar to oneself than they actually are (Ross et al., 1977). After all, if one is competent and they mistakenly think that others are like themselves, then it is unsurprising if they also think others are more competent than they really are—perhaps even more competent than oneself. However, a recent study failed to find support for the idea that high performers overestimated the competence of others in absolute terms, even though they still underestimated their own performance relative to others (Tirso et al., 2019). Consequently, it is unclear as to what explains this paradox.

In any case, the available evidence suggests humans are frequently inaccurate in their metacognition.

4.2 Explanations of Metacognitive Inaccuracy

We might then ask what explains this metacognitive inaccuracy.

Kruger and Dunning claimed that metacognitive inaccuracy is explained by idea that the same things that are necessary for accurate judgments are also necessary for accurate metajudgments. In a highly cited paper, they state:

> In essence, we argue that the skills that engender competence in a particular domain are often the very same skills necessary to evaluate competence in that domain—one's own or anyone else's…. For example, consider the ability to write grammatical English. The skills that enable one to construct a grammatical sentence are the same skills necessary to recognize a grammatical sentence, and thus are the same skills necessary to determine if a grammatical mistake has been made. In short, the same knowledge that underlies the ability to produce correct judgment is also the knowledge that underlies the ability to recognize correct judgment. To lack the former is to be deficient in the latter. (Kruger and Dunning, 1999, pp. 1121–1122)

So that is one possible explanation.

Another possible explanation is that those with inaccurate metacognition lack incentives to be accurate. However, support for this explanation is currently lacking, at least according to my review of the literature. For example, one study tried to improve accuracy by offering students a prize of $50 for the most accurate prediction of their grade, and the prize was repeatedly emphasized to make it salient (Saenz et al., 2019). However, accuracy did not improve despite this incentive. Consequently, it is not clear that inaccuracy always results merely from a lack of motivation to have accurate metacognition.

Another couple of factors may explain metacognitive inaccuracy—or at least the persistence of such inaccuracy.

One factor is that it is not clear that people track the past accuracy of their judgments: there is a lack of record keeping, so to speak. After all, if one has inaccurate judgments and if they fail to pay attention to the track record of their judgments, then it is not surprising that they would fail to realize how inaccurate they really are.

Further support for this comes from studies that report increases in accuracy when participants are given feedback about their past accuracy (Callender et al., 2016; Moore et al., 2017; Saenz et al., 2019). For example, Moore et al. (2017)

report an experimental study in which some participants were given feedback about their accuracy. These participants saw a substantial improvement in their accuracy, mainly by reducing overconfidence. In saying that, it is difficult to know precisely how much of this is attributable to other variables, since their participants and others received instruction about other things aside from just feedback on their past accuracy.

Aside from the lack of record keeping, there is another factor that might explain the persistence of metacognitive inaccuracy: people may have motivational biases *not* to realize how inaccurate they are. As we shall see later, some evidence suggests our motivations may influence the processes by which we seek information and reason from it. For example, Jang (2014) studied the types of information that American adults exposed themselves to online. He found that individuals often exposed themselves to information which contradicted their opinions, but particular individuals were less likely to do this. More specifically, individuals who saw themselves as knowledgeable about science were less likely to view scientific articles purporting to contradict their previously held views. Jang consequently thought this was probably attributed to their desire to avoid information that challenged them. However, the way in which it challenged these people was potentially different: other participants in the study often examined evidence against their opinions, but what made these self-reportedly "knowledgeable" individuals different was *their opinion about themselves*. In particular, individuals who saw themselves as more knowledgeable about science may have been less likely to examine information that undermines their opinions, not because the information undermines their opinions, but rather because the information also undermines their self-esteem and their opinions of how knowledgeable they are.

So it is possible that metacognitive inaccuracy sometimes persists partly because we ignore evidence of inaccuracy due to a desire to preserve self-esteem.

But what's more, some may seek not only to ignore the evidence of their inaccuracy, but to actively distort or dismiss it. For example, recall Tetlock's (2005) study of political judgment. To refresh your memory, he found that many "experts" in political affairs were startingly inaccurate in their judgments. But beyond that, he also found that when some participants were confronted with evidence of their inaccuracy, they protested against the very standards by which their accuracy was assessed. Consequently, they demanded numerous adjustments to the procedures that scored the accuracy of their forecasts about future events. As Tetlock (2005) says:

> We confronted more and more judgment calls on how far to go in accommodating these protests. And we explored more and more adjustments to procedures for scoring the accuracy of experts' forecasts, including *value adjustments* that responded to forecasters' protests that their mistakes were the "right mistakes" given the costs of erring in the other direction; *controversy adjustments* that responded to forecasters' protests that they were really right and our reality checks wrong; *difficulty adjustments* that responded to protests that some forecasters had been dealt tougher tasks than others; and even *fuzzy-set adjustments* that gave forecasters partial credit whenever they claimed that things that did not happen either almost happened or might yet happen. (p. 9)

Ultimately, however, these experts performed poorly by all of these standards. He indicates that he found “the composite statistical portraits of good judgment to be robust across an impressive range of scoring adjustments” (p. 9). If anything, this served to reinforce Tetlock’s confidence in the overconfidence of the experts' judgments.

But beyond that, this leaves one with the impression that those who would be most devasted by evidence of their inaccuracy will also go to great lengths to dismiss it.

4.3 Summary

In sum, then, not only are humans often inaccurate in their judgments, but they are also often inaccurate in their metacognition about those judgments.

There are various explanations of why this is the case.

One explanation is that humans lack the motivation to have accurate self-assessments, but there is little (if any) evidence currently supporting this explanation. Instead, some evidence found that particular incentives failed to reduce inaccuracy.

Regardless, other evidence suggests that two factors may explain why this megacognitive inaccuracy persists: first, people often do not track the accuracy of judgments, and furthermore, they are susceptible to dismiss evidence of their inaccuracy to the extent that such evidence can compromise their self-esteem. In a sense, motivation might not suffice for accurate metacognition (since track records or other things may be necessary too), but motivation may suffice for inaccurate metacognition (since, for example, it can lead one to dismiss track records).

However, while these factors may explain why this inaccuracy persists, they might not explain why it arises in the first place.

Kruger and Dunning (1999) offered an explanation: the abilities that make us competent in a domain are also the skills that are necessary to recognize when we are incompetent in that domain. In other words, we need to the ability to form accurate judgments in order to accurately judge how accurate those judgments really are.

This explanation may be true, but even so, it does not really explain how people do arrive at their judgments, even if they do so in ways that do not conduce to accuracy.

So, we might then wonder: how do we arrive at our judgments?

That is the subject of the next chapter. There, we will see that not only are people often unaware of the accuracy of their judgments, but if dual-systems theory is anything to go by, then they are often somewhat unaware of the unreliable ways in which they arrive at their judgments in the first place.

References

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*(1), 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002

Hall, S. R., Stephens, J. R., Seaby, E. G., Andrade, M. G., Lowry, A. F., Parton, W. J. C., Smith, C. F., & Border, S. (2016). Can medical students accurately predict their learning? A study comparing perceived and actual performance in neuroanatomy. *Anatomical Sciences Education, 9*(5), 488–495. https://doi.org/10.1002/ase.1601

Händel, M., & Dresel, M. (2018). Confidence in performance judgment accuracy: The unskilled and unaware effect revisited. *Metacognition and Learning, 13*(3), 265–285. https://doi.org/10.1007/s11409-018-9185-6

Jang, S. M. (2014). Seeking congruency or Incongruency online? *Science Communication, 36*(2), 143–167. https://doi.org/10.1177/1075547013502733

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

Lechuga, J., & Wiebe, J. S. (2011). Culture and probability judgment accuracy: The influence of holistic reasoning. *Journal of Cross-Cultural Psychology, 42*(6), 1054–1065. https://doi.org/10.1177/0022022111407914

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning Memory and Cognition, 37*(2), 502–506. https://doi.org/10.1037/a0021802

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science, 63*(11), 3552. https://doi.org/10.1287/mnsc.2016.2525

Ross, L., Greene, D., & House, P. (1977). The "false in social consensus perception effect": An egocentric bias and attribution processes. In. *Journal of Experimental Social Psychology, 13, Issue 3*. https://doi.org/10.1016/0022-1031(77)90049-X

Saenz, G. D., Geraci, L., & Tirso, R. (2019). Improving metacognition: A comparison of interventions. *Applied Cognitive Psychology, 33*(5), 918–929. https://doi.org/10.1002/acp.3556

Tetlock, P. (2005). Expert political judgment: How good is it? How can we know? Princeton University Press.

Tirso, R., Geraci, L., & Saenz, G. D. (2019). Examining underconfidence among high-performing students: A test of the false consensus hypothesis. *Journal of Applied Research in Memory and Cognition, 8*(2), 154–165. https://doi.org/10.1016/j.jarmac.2019.04.003

Chapter 5
How We Think: The Rationality of Our Reasoning

We have seen that people are often inaccurate, both in their judgments, as well as in their metajudgments about the accuracy of those judgments.

But how do people get to be so inaccurate in the first place?

To some extent, we can explain the inaccuracy of our judgments by examining the processes that formed those judgments. For the most part, I refer to such judgment forming processes as our *reasoning* (although note that others, such as Mercier and Sperber (2017), use the term "reasoning" in a more restricted sense).

5.1 Rationality, Heuristics, and Biases

Philosophy and psychology are especially concerned with standards by which to assess the *rationality* of our reasoning.

But what does "rationality" mean in this context?

Well, there are different understandings of rationality. Philosophers commonly distinguish between *epistemic* or *theoretical rationality* and *practical* or *pragmatic rationality* (Audi, 2003). In Robert Audi's (2003) words, epistemic rationality has to do with "the rationality of cognitions, such as beliefs, in virtue of which we are theorizing beings seeking a true picture of our world" (p. 18). In contrast, pragmatic rationality has to do with "the rationality of elements, such as actions, in virtue of which we are practical beings seeking to *do* things, in particular to satisfy our needs and desires" (Audi, 2003, p. 18). Furthermore, pragmatic rationality is sometimes also understood in two different ways: *procedural rationality*, which is a matter of having desires that are in some sense implied by one's other desires and judgments, and *substantive rationality*, which is a matter of having desires that are in some sense the right desires to have, regardless of whether those desires are implied by one's other desires and judgments (Hooker & Streumer, 2004).

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0_5

Not all of these understandings are incompatible: one might think, as Audi (2003) does, that there are two distinct types of rationality, one epistemic and one pragmatic, both of which are interconnected. For example, if we are to pragmatically act in ways that satisfy our desires, then we often need epistemically rational judgments about what will satisfy those desires.

In any case, there are different understandings of rationality, and not all of them are described in exactly the same ways. Stanovich et al. (2016), for instance, describes epistemic rationality in terms of how well our beliefs "map onto the actual structure of the world" (p. 6).

Here, we will focus on the *epistemic rationality* of our judgment forming processes, loosely understood as the extent to which such processes conduce to accurate judgments.

In the study of reasoning, judgments are typically assessed by their conformity to the standards of logic and probability theory (Chater & Oaksford, 2012).

As it turns out, however, people's thinking processes often depart from these standards, and the heuristics and biases research program has uncovered a wealth of insight into why this is. Here, the term "bias" refers to a deviation from a normative standard of rationality (Gilovich & Griffin, 2002). A "heuristic" then refers to a kind of thinking process—one which can explain how the bias was produced. Kahneman states that, "[t]he technical definition of *heuristic* is a simple procedure that helps to find adequate, though often imperfect, answers to difficult questions" (Kahneman, 2011, p. 98).

5.2 Dual-Process Theory: System 1 and System 2

The heuristics are often described in the context of *dual-process theory*. Dual process theory is extremely popular, both in and outside of cognitive psychology. Despite this, some scholars disavow dual-process theory, such as Mercier and Sperber (2017) who claim (rightly or wrongly) that dual-process theory is on the decline in cognitive psychology.

Dual-process theories describe psychological phenomena—such as reasoning, emotion regulation, or other things—in terms of two kinds of processes. Each of these processes are said to be carried out by a different "system" in the mind. The labels and precise characterizations of these processes often differ: in fact, Stanovich (2012) identified 27 (!) distinct ways that scholars have tried to distinguish these processes in the literature. Regardless, we will follow Kahneman and Stanovich in referring to these two processes with the imaginative labels "Type 1" and "Type 2" processes, each of which are carried out by "System 1" and "System 2," respectively.

We will also follow Stanovich's account of these systems. Stanovich et al. (2016) call the reasoning of System 1 "Type 1 processing". According to them, the "defining feature of Type 1 processing is its autonomy" (Stanovich et al., 2016, p. 17). By that, they mean that such processes "execute automatically upon encountering their triggering stimuli, and they are not dependent on input from high-level control systems" (Stanovich et al., 2016, p. 17). (We will see shortly what such a "high level control system" amounts to.) Type 1 processes typically have other features: they are rapid, they are not burdensome on one's cognitive faculties, and they tend to be associative (i.e., they are often a matter of making simple associations between one thing and another thing). Regardless, Stanovich et al. (2016) claim that these features are not essential to Type 1 processing; they are just correlated with it.

Kahneman (2011) uses a perceptual example of a Type 1 process: the visual recognition of facial emotions. We can quite often automatically recognize people's emotions from their faces: a happy face immediately looks happy, and a sad face immediately looks sad. There is often little cognitive processing to determine what emotions are expressed by someone's face.

Contrast this to System 2 thinking. Again, this is characterized differently from author to author. But it typically has the following features: it is relatively slow, it is cognitively burdensome in the sense that it is often difficult and requires relatively heavy use of cognitive faculties like attention and memory, and it is serial in the sense that it focuses on one task at a time, rather than Type 1 processes which, in theory, can operate rapidly in parallel (Stanovich et al., 2016).

Kahneman uses a mathematical example to illustrate Type 2 thinking: for example, try to multiple 17 by 24. If you do, you will find that the thinking process is very different to when we recognize emotions from faces. Instead, the multiplication process is much slower and cognitively burdensome. That, then, is a System 2 thinking process.

Stanovich further characterizes System 2 in terms of two, if you like, subsystems: the algorithmic mind and the reflective mind. The "algorithmic mind" refers to *cognitive abilities* when reasoning. This includes, for example, the cognitive abilities at play when recalling information about the frequency of something (like the frequency of shark attacks) and using that to estimate the probability of that thing.

The "reflective mind" refers to *thinking dispositions*—or "cognitive styles" as they are sometimes called. Examples of thinking dispositions include the tendencies to gather more evidence before reaching a conclusion, to consider alternative viewpoints, to consider both the pros and the cons of a proposed course of action, and so forth. The reflective mind is then a "high level control system" which initiates the override of conclusions from Type 1 processes, while the algorithmic mind carries out the processes that are necessary to replace the conclusions of such Type 1 processes with better ones.

Of course, I follow others in using the labels "system" and "mind," but I do not claim that the mind is in some sense metaphysically complex, thereby comprised of different parts with different labels. Maybe such a claim is true; maybe it is not. Rather, the labels are used simply to denote different types of thinking processes, regardless of the metaphysics of the thing that carries out the processes. I have simply followed others in using their labels to refer to these processes.

One's thinking can then be evaluated in terms of these processes. Rationality and good reasoning are often then characterized in terms of having specific thinking dispositions and cognitive abilities, like those mentioned earlier. Practically, this is largely measured by tasks that prompt incorrect Type 1 reasoning. Good reasoners, it is thought, are those who override such Type 1 processes with Type 2 reasoning involving such dispositions and abilities.

The heuristics have typically been studied through tasks that prompt biased Type 1 reasoning, since such biases are thought to provide insight into the operations of the mind and these heuristics.

5.3 Misconceptions About Heuristics and Type 1 Processing

However, there are various potential misconceptions about these heuristics and systems.

The first is about whether the biases and heuristics are always the automatic products of lazy minds or whether they can also result from controlled and deliberate thinking. Some refer to quick, automatic, and lazy judgments as the judgments of a "cognitive miser"—someone who expends as little cognitive effort as possible (Stanovich et al., 2016). One might then think that such biases and heuristics result purely from cognitive miserliness—so to speak.

However, Tversky and Kahneman (1983) explicitly saw heuristics as being strategies that may or may not be used deliberately. Further support for this perspective comes from a study of Camerer et al. (1999). They reviewed 74 experiments and found that financial incentives sometimes improved performance, but not always: in fact, often they do not. Instead, they claim that "*no* replicated study has made rationality violations disappear purely by raising incentives" (Camerer et al., 1999). Gilovich and Griffin (2002) suggest such findings show that heuristic thinking and biases are not always attributable to cognitive miserliness, even if they often are. Instead, such biases are often understood as *cognitive illusions,* illusions that are often difficult to detect and to override since System 2 may have "no clue to the error" (Kahneman, 2011, p. 28). This, for what it's worth, also resonates with my own experimental research on biases which have not disappeared merely by raising incentives; such biases often look more like cognitive illusions than products of lazy thinking.

A second misconception is that heuristics and Type 1 processes are always irrational in the sense that they never conduce to truth. This, however, is not the case.

Sometimes they produce efficient judgments that are accurate enough for the context. This can occur if the environment is said to be "benign"—that is, containing useful cues that can be utilized by Type 1 processes to produce accurate judgments or adaptive behavior (Stanovich et al., 2016). The study of *ecological rationality* refers to the study of the environments where a heuristic will "succeed" and those environments where it will "fail," to use the words of Gigerenzer and Brighton (2009). Gigerenzer and Brighton (2009) also argue that there are some environments where heuristics perform better than other ways of thinking. These environments exhibit a *less-is-more* effect—a phenomenon where less information, time, and processing can *improve* accuracy. Type 1 thinking can also be adequate in contexts where proficiency in a domain is practiced and the relevant judgmental processes become both automatic and accurate (Kahneman, 2011).

However, it is clear that heuristics and Type 1 processes often produce inaccurate judgments. This is especially the case when environments are said to be "hostile" (Stanovich et al., 2016). Such hostile environments are those that contain insufficient cues that are misleading or other agents who will try to exploit Type 1 processes (as in the case of manipulative advertising).

5.4 Search Heuristics and Inference Heuristics

So we have discussed dual-process theory, as well as the heuristics which influence how people arrive at their conclusions.

It will be useful to further divide these heuristics in two kinds: search heuristics and inference heuristics. A "search heuristic" refers to a thinking process by which we *seek* out or *search* for information. An "inference heuristic" refers to a thinking process by which we *arrive* at conclusions on the basis of the information we have. The former kind refers to how we find facts and the other to how we use those facts to arrive at conclusions.

Heuristics of both sorts can explain how humans arrive at inaccurate judgments, and we will now review some of these.

5.4.1 Motivation, Search Heuristics, and Confirmation Bias

It is clear that our motivations influence how we arrive at our conclusions. This is not always in a bad way: we might be motivated to examine as much evidence as possible and to examine it in as fair a way as possible, for instance.

More generally, then, the psychology of reasoning distinguishes several ways in which motivations can influence how we arrive at our conclusions.

One of these concerns the distinction between *outcomes* and *strategies*: one might be motivated to reach a specific *outcome* of their investigation, or they might be motivated to employ a specific *strategy* in their investigation, regardless of the strategy's outcome (Molden & Higgins, 2012).

Furthermore, another distinction concerns the types of outcomes one might desire: the distinction between *directional* and *nondirectional* outcomes. One desires a "directional" outcome when they are motivated to think in ways that lead them to a specific conclusion. Various motives have been studied in this respect, including the motive to reach specific conclusions which enhance or preserve things like self-esteem (Sanitioso et al., 1990), like social connection with others (Molden & Maner, 2013), or like feelings of control and meaning (Heine et al., 2006). Examples of such conclusions may be, for example, the conclusions that you are intelligent, that your opinions are not mistaken, that you are similar to someone who you are romantically interested in, and the like.

In contrast, one desires a “nondirectional” outcome when they are motivated to think in ways that lead them to a conclusion with specific attributes, regardless of what that particular conclusion is. One might be motivated, for instance, by the desire to reach a conclusion that is accurate (Neuberg & Fiske, 1987) or that satisfies one’s need for closure (Kruglanski & Webster, 1996). Such attributes are, in principle, compatible with a wide range of specific conclusions; an accuracy motivation can lead you to the conclusion that you are intelligent or to the opposite, depending on which conclusion you ultimately think is more accurate.

Of particular importance is one search heuristic: seeking out information that is compatible with your opinions (or your desired opinions). This is also called *confirmation bias, selective exposure, myside bias* or *congeniality bias* (Hart et al., 2009).

Some studies are consistent with the possibility that humans exhibit confirmation bias.

For example, Jang (2014) conducted a study with a demographically diverse sample of 238 American adults. Participants filled out a survey about their personal views and were then free to browse science articles on a science news website. He tracked what kinds of information participants exposed themselves to.

Participants generally exposed themselves to information that *challenged* their previous views on topics such as stem cell research and genetically modified foods. He states his findings suggest that “online users may not be as susceptible to confirmation bias” as some researchers think (Jang, 2014, p. 159). To that extent, many appeared free of confirmation bias.

Others, however, did not. In particular, he found that two groups of people were less likely to view information contradicting their opinions: those who thought they had a lot of knowledge about science already and those who were highly religious. Religious people, for instance, had no aversion to reading articles supporting their views on topics like stem cell research and evolution. However, they were significantly less likely to read science news articles *against* such views. Jang also claims that those who thought they knew a lot of science and felt confident in their opinion probably “would not want to confront challenging information that may result in cognitive dissonance” (Jang, 2014, p. 160).

There is, however, a difficulty in interpreting studies like these: sometimes it is rational to let prior beliefs discount conflicting evidence. For example, suppose you came across an opinion piece from a scientist who claimed to have produced evidence that the earth is completely flat instead of being round. You might quite rightly think this opinion piece undermines the credibility of the scientist more than your views about whether the earth is round, and so you might decide not to look at the opinion piece for that reason. In this case, your prior beliefs might be based on very, very good evidence, so much so that it merits discounting conflicting evidence like the scientist’s opinion. For that reason, it is difficult to determine whether selective exposure to conflicting evidence is irrational or not. This idea has come to be associated with what is called the *knowledge projection* argument in the psychological literature (Stanovich, 2021).[1]

In any case, Jang's (2014) study is certainly consistent with the possibility that some people exhibit pernicious confirmation bias when searching through evidence.

There are broader reviews of studies about confirmation bias. A widely discussed classic is from Kunda (1990). At the time she wrote her article, she argued that the available evidence was consistent with a specific account of motivated reasoning. On her account, one's motivation for a conclusion to be true may lead themselves to pose so-called *directional questions*. For example, they may be motivated to think that they are healthy, and so they ask the question "Am I healthy?" instead of "Am I ill?". Or they may be motivated to think they will win a competition, and so they ask the question "Will I win this competition?" instead of "Will I lose this competition?". Once they pose a directional question, they then selectively search for evidence that confirms the affirmative answers, but not directly because they desire those answers to be true. Instead, the influence of motivation is indirect: their motivation leads them to pose the question in a particular way, and then they seek evidence for an affirmative answer merely because humans have a greater tendency to seek confirming evidence for hypotheses rather than disconfirming evidence (regardless of whether they want the hypotheses to be true or not). Of course, Kunda stated that other accounts of the role a motivation are possibly true. Nevertheless, she said that "[i]t is difficult to tell, at this point, whether the effects of motivation of reasoning go beyond the posing of directional questions" (Kunda, 1990, p. 495).

While Kunda's classic piece helpfully charts out some theories about motivation, more comprehensive reviews of the literature have since been published.

One of these is a meta-analysis that examines confirmation bias across 67 articles describing 91 studies with just under 8,000 participants (Hart et al., 2009). The authors report that the studies reveal a general confirmation bias in their participants: that is, a preference for exposure to information that accorded with their views. However, they state that the bias was not always present, and it was influenced by several variables. For instance, confirmation bias was more likely among participants when their attitudes, judgments, and behaviors were relevant to their values or were held with conviction. It was also more likely when participants scored high on measures of closed-mindedness, as assessed by questions about, for example, their openness toward people who think differently.

However, this meta-analysis has drawn some criticism. Hahn and Harris (2014), for instance, claim that it conflates the desire for information that is consistent with one's *beliefs* with the desire for information that is consistent with their *likings*—or, as psychologists say, their *attitudes*. They claim that the former is an irrational form of motivated reasoning whereas the latter is not necessarily irrational, since there are objective standards for evaluating one's beliefs as true but no similarly objective standards for evaluating one's likings.

[1] Stanovich (2021) also provides an excellent overview of the literature on myside bias. But that said, it focuses more on *distal beliefs* or *convictions* which concern ethical topics, like whether abortion or guns *should* be banned. For that reason, Stanovich's excellent work is beyond the purview of this book, which, as mentioned in the introduction, concerns more *descriptive* judgments than *ethical* ones.

In any case, it is plausible that people seek out information in accordance with their motivations. The important point is that these motivations may sometimes deter individuals from accuracy by, for instance, deterring them from counter-evidence to their judgments. To that extent, the motivation to find evidence consistent with one's judgments can be regarded as a search heuristic that compromises accuracy.

Of course, this is just one search heuristic, and there may be others. But the topic of search procedures has received less attention in the literature than the other kind of heuristic: inferential heuristics, processes by which we arrive at our judgments on the basis of the information we have. We will consider some of these now.

5.4.2 Availability Heuristic

One important heuristic is the *availability heuristic*. The heuristic has been described differently by different authors (Higgins, 1996; Tversky & Kahneman, 1973), and some use terms like "retrieval fluency" to refer to it or similar concepts (Schooler & Hertwig, 2005).

Regardless, we can understand the heuristic as such. Suppose one is estimating a probability or a frequency: for example, the probability that a randomly selected individual died by homicide, or the frequency with which homicides occur. Then, the availability heuristic is the process whereby probabilities or frequencies are estimated based on the mental "availability" of relevant instances—that is, the ease with which relevant instances can be called to mind.

One study examined participants' perception of the frequency with which people die as a result of various causes (Lichtenstein et al., 1978). They report that people overestimated "sensational" causes of death, such as tornadoes, homicides, and car accidents (Lichtenstein et al., 1978). For instance, the study reports that participants thought that more people died from floods than from asthma, even though the reverse is true: at the time, asthma claimed nine times as many lives as floods. They further report that measures of availability did a good job of predicting participants' overestimation, and they claimed availability appeared relevant to explaining their findings.

Some think that media coverage is in turn responsible for what things are more available than others in the public consciousness. Reber (2017) claims that violent crimes are frequently covered in the media and are consequently more mentally available, thus leading to overestimation of the frequency of violent crimes.

Schwarz and his colleagues have also examined whether this kind of inferential bias results from either of two causes: 1) the *ease* with which instances can be called to mind or 2) the *amount* of instances that can be called to mind. For example, in a series of studies, they asked participants to rate their own assertiveness or unassertiveness (Schwarz et al., 1991). Prior to this, some participants were asked to recall 12 instances where they were assertive, and others were asked to recall 6 such instances. Those who tried to recall only 6 instances rated themselves as more assertive compared to those who tried to recall 12. They claimed that participants appeared to conclude that "they can't be that assertive (or unassertive) if it is so difficult to recall the requested number of examples" (Schwarz et al., 1991, p. 201).

It seems that, according to them, the inferential influence of availability is mediated by "second-order" beliefs about the availability: participants consciously believe that high availability supports high probability or frequency estimates and low availability should reduce these estimates. That, then, is Schwarz et al.'s take, although I will reconsider this take shortly.

In any case, the availability heuristic often influences judgments, but not always. Some evidence suggests its influence is either reduced or absent in two contexts.

One context concerns cases where the topic is of high personal importance. In two sets of studies, individuals were asked to estimate their risk of experiencing something bad—sexual assault in one set of studies (Grayson & Schwarz, 1999) or heart disease in the other (Rotliman & Schwarz, 1998). The studies' authors reported that, on average, availability played a reduced role for those who thought that the risks were especially relevant to them: these were people who believed that anyone like themselves one could experience sexual assault in the former study and people who had a family history of heart disease in the other. For these people, they were reportedly more likely to rely on the specific content of what they could recall, rather than the availability of what they could recall.

A second context concerns when the participants consciously believe that availability is *not* diagnostic: that is, the participants believe that their mental availability is not informative about the frequency or probability they are estimating. In two sets of studies, for instance, availability reportedly played a reduced role when participants were told that some background music would inhibit the ease with which they could recall relevant instances (Haddock et al., 1999; Schwarz et al., 1991). The idea was that participants attributed their difficulties in recollection to the music, which they heard, and so they would not treat this difficulty as informative about the frequency or probability, which they were estimating.

Again, from these studies, it appears that the influence of availability is mediated by conscious beliefs about the significance of availability.

But in saying that, it is not obvious that this is the final word.

In particular, it is not obvious that the inferential impact of availability is always mediated by conscious beliefs about the significance of availability, even though it may be mediated as such in Schwarz's experiments. Instead, it seems quite plausible that the inferential impact of availability could frequently be more automatic: that is, availability somewhat influences probability and frequency estimates without the participants having conscious beliefs about it. If it is automatic in some conditions, then Schwarz's experiments may provide evidence of other conditions where conscious beliefs can interact with availability to influence its inferential impact: in particular, participants can reduce probability and frequency estimates on the basis of conscious beliefs about a lack of availability, and the influence of these conscious beliefs can in turn diminish when participants also believe that this lack of availability is not significant.

Some studies also suggested that the vividness of recalled information can affect judgment, either by making the information more available (Reyes et al., 1980) or possibly by other means (Blondé & Girandola, 2018; Shedler & Manis, 1986). Regardless, other studies have failed to find evidence for the influence of vividness (Bell & Loftus, 1988; Collins et al., 1988; Rogers et al., 2016). Consequently, it is still unclear about under what conditions, if any, vividness influences judgments.

5.4.3 Representativeness Heuristic

Another widely studied heuristic is known as the representativeness heuristic.

The most famous illustration of its application is the case of Linda (Tversky & Kahneman, 1982, 2002). In one study, a group of 88 students were given the following description of Linda:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations. (Tversky & Kahneman, 1982. p. 92)

They then had to rank the probability of various statements about Linda. Interestingly, participants thought it was more probable that "Linda is a bank teller and is active in the feminist movement" than that "Linda is a bank teller". But clearly this is erroneous: it cannot be the case that Linda is more likely to be *a feminist* bank teller than that she is a bank teller who may or may not be a feminist. After all, the set of all bank tellers *includes* the set of all bank tellers who are feminists, and one cannot be more likely to be in a subset of all bank tellers than in the set of bank tellers themselves. The participants' rankings then exhibited *the conjunction fallacy*, a fallacy whereby a conjunction of propositions is regarded as more probable than one of its conjuncts.

Tversky and Kahneman (1982) proposed that this judgment was made because Linda seems more *representative* of the category of feminist bank tellers compared to the category of bank tellers in general. The idea is that participants consider the information about Linda and then compare her to stereotypical instances of both feminist bank tellers and bank tellers more generally. The stereotypical feminist bank teller is presumably a woman who cares about social justice. But the stereotypical bank teller is not necessarily a feminist, nor a woman, nor someone who cares about social justice. However, since Linda is a woman who, among other things, is concerned about social justice, she seems more similar to a stereotypical feminist bank teller than to a stereotypical bank teller. According to the representativeness heuristic, the probability that she belongs in either category is then estimated on the basis of how similar she is to the stereotypical instances of either category. Consequently, the representativeness heuristic can explain why participants thought Linda is more likely to be a feminist bank teller than a bank teller in general.

Note that availability may also play a role here. In particular, the stereotypical instance of a category may be nothing more than the most "available" instance of that category which comes to mind.

In any case, there have been numerous challenges to the existence of this bias, but this bias in participants' judgment has also persisted under various modifications of the experiment. For instance, in a later study, participants used a rating scale to rate the probability these two statements:

T&F: Linda is a bank teller and is active in the feminist movement.
T*: Linda is a bank teller whether or not she is active in the feminist movement.

Participants on average still rated the former statement (T&F) as more probable than the latter statement (T*), and this is despite that fact that the latter statement explicitly involved the possibility that Linda was a feminist bank teller (Tversky & Kahneman, 2002).

The Linda case, then, is the most well-known illustration of representativeness, a well-replicated phenomenon.

More generally, though, representativeness refers to when someone or something's membership in a category is estimated based on how similar that person or thing is to the most typical or available instance of that category.[2]

Representativeness has also been used to explain numerous other biases, including the biases known as "base-rate neglect", the "gambler's fallacy", and belief in the so-called "law of small numbers" (Teigen, 2017).

5.4.4 *Anchoring Heuristic*

Another heuristic is known as the *anchoring heuristic*, or the *anchor and adjustment effect.*

An early example of anchoring is described by Tversky and Kahneman (1974). Participants were asked to estimate the percentage of African countries in the United Nations. To do this, the experimenters spun a wheel of fortune in front of the participants, and the wheel randomly selected a number between 0 and 100. This number is what would be called the *anchor*—the number or the initial estimate presented to a participant for their consideration. Participants were asked to indicate whether the percentage of African countries in the United Nations was greater or less than the anchor and to then to estimate the true percentage by moving up or down from that anchor. The results found that the arbitrary anchors affected the estimates of participants: the group with an anchor of 10 had a median estimate of 25 and the group

[2] Tversky and Kahneman (1982) define representativeness as "a relation between a process or a model, M, and some instance or event, X, associated with that model" (p. 85). They do not precisely define what they mean by "a model," but they say such models "could be of a person, a fair coin, or the world economy," while the things that are associated with that model might include "a comment, a sequence of heads and tails, or the present price of gold" (Tversky & Kahneman, 1982, p. 85). Aside from that, I found it difficult to get clearer on precisely what they think the heuristic involves.

with an anchor of 65 had a median estimate of 45. The use of incentives also failed to reduce this effect.

Since then, the anchoring heuristic has been found in a variety of other contexts, including estimates of risk and probability (Wright & Anderson, 1989), willingness to pay for consumer products (Ariely et al., 2003), and one's evaluation of their ability to meet the requirements of a given situation (Cervone & Peake, 1986).

It is not clear as to precisely what explains the anchoring effect, even though various explanations are potential candidates.

One explanation involves *selective accessibility*—the phenomenon where the anchor makes select information more accessible than other information (Mussweiler & Strack, 2000). Mussweiler and Strack (2000) described some studies in support of this explanation. However, a later replication attempt failed to replicate their findings (Harris et al., 2019), even though the replication still revealed evidence of anchoring. Consequently, it is unclear as to whether this explanation is true.

A second explanation is that some effects are just a result of a "basic anchoring" effect, without regard to any increased accessibility of information (Critcher & Gilovich, 2008). Critcher and Gilovich (2008) described some studies in support of this explanation. But again, a subsequent replication attempt failed to replicate their findings (Open Science Collective, 2018), and so it is also unclear as to whether this explanation is true.

A third explanation is *scale distortion theory*. Frederick and Mochon (2011) describe this theory with an analogy. They claim that the teeth of someone with dark skin appears whiter than the teeth of someone with whiter skin. They think this illustrates that our perception of something is influenced by the contextual considerations which it is compared with. In scale distortion theory, however, the contextual considerations do not influence one's perception of the thing they are estimating, such as the quantity of African countries in the United Nations. Instead, the contextual considerations influence the one's perception of the *scale* used to communicate that thing. Frederick and Mochon's (2011) idea is no different: "the perceived magnitude of a number is affected by other numeric values on that scale with which it is compared. For example, 900 pounds seems larger if compared with 20 pounds and seems smaller if compared with 5,000 pounds" (p. 1). Consequently, one might have an impression of something's weight, but use a different number to communicate that same impression depending on whether an anchor makes the number look bigger or smaller. They likewise describe some studies in support of their theory (Frederick & Mochon, 2011; Mochon & Frederick, 2013), but, so far as I can tell, these are yet to be replicated by other researchers.

In any case, evidence for the anchoring effect has been replicated (Furnham et al., 2012; Jung et al., 2016; Marcus & Philipp, 2017; Teovanović, 2019), even though some evidence for particular explanations of it has not.

5.4.5 Motivated Reasoning

Earlier, we saw that motivation might affect the way in which individuals search for information to inform their conclusions.

However, motivation may also play a role in inference—that is, a role in how people arrive at their conclusions on the basis of information they already have. (This is sometimes called “selective interpretation” of information, as opposed to “selective exposure” to information.) In two studies, for example, a combined total of 96 heterosexual undergraduates viewed profiles of people of the opposite sex (Slotter & Gardner, 2009). In each study, the undergraduates were split into two groups: the experimental group and the control condition. The experimental group saw profiles of fellow students of the opposite sex on a dating website. The control group saw essentially the same profiles, but on a non-romantic website featuring the profiles of fellow students running for student government offices. The participants were then asked to rate their similarity to the individuals in the profiles. They report that participants in the experimental group rated themselves as more similar to the individuals in the profiles on average than the participants in the control group. This was the case even for qualities that participants had previously indicated were *not* characteristic of themselves two weeks earlier. The study authors attribute this difference in ratings to differences in motivations: those in the experimental group were more motivated to see themselves as similar to the individuals in the profiles since it was in their romantic interest to see themselves as similar.

That said, there have not yet been published attempts to replicate these findings, so they must be interpreted with caution. And even then, it is not clear whether participants *believed* they were more similar in the experimental condition or instead whether they merely *reported* that they were more similar, perhaps because they thought doing so might affect their chances of interacting with the potential romantic interests.

Regardless, these studies at least suggest that it is possible that motivation can affect not only how people search out information but also how they reason from information they already have. After all, the participants in the conditions did not differ in terms of the information they had about their similarity to those in the profiles. They differed only in that they viewed the profiles in different contexts. These contexts may have sufficed to motivate individuals to recall or interpret their available evidence in ways that led them to desirable conclusions.

5.5 Social Influences

So we have surveyed a variety of processes through which we arrive at our judgments.

However, to a large extent, we have focused mainly on processes at the *individual* level: for instance, availability, representativeness, and other heuristics are processes that take place in an individual’s mind.

Yet we all are embedded in societies that influence the way we arrive at our beliefs. While an exhaustive examination of this influence is impossible here, we can at least consider some sources of social influence.

Philosopher of science Cailin O'Connor and physicist James Weatherall (2019) identify various social factors that explain the development or persistence of false beliefs.

One of them is *conformity*—or, in this case, specifically the phenomenon of conforming one's beliefs to those of their peers. O'connor and Weatherall cite the famous and widely replicated studies of Solomon Asch (Asch, 1951; Larsen, 1990; Ušto et al., 2019). In these studies, a participant gives an obviously incorrect answer to an easy question merely because they are surrounded by actors who act as genuine participants but who intentionally affirm the obviously incorrect answer. O'connor and Weatherall think this phenomenon can explain the persistence of false beliefs in communities, such as anti-vaccination attitudes in the Somali community of Minnesota. The idea is that if many members of a community come to suddenly accept a belief, conformity can ossify the belief and make it resistant to change or counter-evidence. Furthermore, O'connor and her colleagues conducted computer simulations to model how conformity affects the ability of the scientific community to reach the truth and to form consensus about it. Her results indicated that conformity inhibited the spread of good new ideas among scientists. Furthermore, she found that "on average, the greater their tendencies to conform, the more often a group of scientists will take the worse action. When they care only about performing the best action, they converge to the truth most of the time" (O'Connor & Weatherall, 2019, p. 86).

Another social factor that explains the development and persistence of false beliefs is what philosophers of science Bennett Holman and Justin Bruner call *industrial selection* (Holman & Bruner, 2017). Industrial selection occurs when companies fund research groups that already use methods that favor the interests of the companies. The funding in turn helps the research groups acquire resources and positions of prominence so that their ideas can spread. In essence, it is when companies fund and support the proliferation of ideas which profit them. In this way, it resembles evolution; natural selection favors species that are evolutionarily fit, while industrial selection favors ideas that are profitable. O'connor and Weatherall (and Holman and Bruner) claim that industrial selection explained the proliferation of antiarrhythmic drugs, that is, drugs that prevented irregular heartbeats. In particular, they claim corporations funded researchers using methods that found that antiarrhythmic drugs prevented irregular heartbeats. Because these researchers already employed these methods prior to funding, however, they could claim that industry did not in any way bias their research. Yet because of the funding, their ideas proliferated. The public and many institutions then came to believe in the safety of these drugs. Unfortunately, however, the drugs actually *caused* heart attacks, even though they prevented irregular heartbeats. As a result, not only did industrial selection cause false beliefs to spread, but it resulted in a false confidence in the drugs which killed many Americans. And indeed, one writer goes as far as placing the death toll in the tens of thousands (Moore, 1995).

Of course, while this is far from a comprehensive review of the social influences of belief, it served to highlight the role that such influences can play, sometimes perniciously.

5.6 Summary

This chapter considered how we arrive at inaccurate judgments by considering the rationality of the reasoning processes by which we arrive at our judgments in the first place. In this chapter, the focus was on *epistemic rationality*, which concerns the accuracy of how we arrive at our judgments.

Dual-process theory states that people arrive at their judgments through either one of two kinds of processes: Type 1 processes, which are fast, intuitive, and often less reliable, and Type 2 processes, which are slower, more deliberate, and often more reliable.

People employ a range of heuristics through either process, heuristics which can explain how they arrive at their judgments. This chapter discussed the most widely known heuristics, including the representativeness heuristic (where the probability that something belongs to a category depends on the extent to which it appears representative of that category), the availability heuristic (where the probability or frequency of something is estimated based on how easily it comes to mind), and others. There is strong replicable evidence for the operations of each of them. There are also other kinds of heuristics discussed by others (e.g. Gigerenzer & Brighton, 2009).

These biases and heuristics are not always the result of lazy or unmotivated minds, nor are they always irrational or inaccurate. They can result from more deliberate thinking too, and sometimes they produce rational and accurate judgments.

Aside from the heuristics, social factors can influence how we arrive at our judgments as well, including conformity and industrial selection.

Ultimately, then, these heuristics and social factors can often explain how we arrive at inaccurate or epistemically irrational judgments.

By why would we have minds that are capable of such sub-optimal heuristic thinking in the first place? That is the topic of the next chapter on our evolutionary origins.

References

Ariely, D., Loewenstein, G., & Prelec, D. (2003). “Coherent Arbitrariness”: Stable demand curves without stable preferences. *The Quarterly Journal of Economics, 118*(1), 73–106. https://doi.org/10.1162/00335530360535153

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. *Organizational Influence Processes, 58*, 295–303.

Audi, R. (2003). Theoretical rationality: Its sources, structure, and scope. In A. R. Mele & P. Rawling (Eds.), *Oxford Handbook of Rationality* (pp. 17–44). Oxford University Press.

Bell, B. E., & Loftus, E. F. (1988). Degree of detail of eyewitness testimony and mock juror judgments. *Journal of Applied Social Psychology, 18*(14), 1171–1192. https://doi.org/10.1111/j.1559-1816.1988.tb01200.x

Blondé, J., & Girandola, F. (2018). Are Vivid (Vs. Pallid) threats persuasive? Examining the effects of threat vividness in health communications. *Basic and Applied Social Psychology, 40*(1), 36–48. https://doi.org/10.1080/01973533.2017.1412969

Camerer, C. F., Hogarth, R. M., Booth, W. W., & Science, B. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. In *Journal of Risk and Uncertainty* (Vol. 19, Issue 13). Kluwer Academic Publishers.

Cervone, D., & Peake, P. K. (1986). Anchoring, efficacy, and action. The influence of judgmental heuristics on self-efficacy judgments and behavior. *Journal of Personality and Social Psychology, 50*(3), 492–501. https://doi.org/10.1037/0022-3514.50.3.492

Chater, N., & Oaksford, M. (2012). Normative systems: Logic, probability and rational choice. In *The Oxford handbook of thinking and reasoning* (pp. 11–21). Oxford University Press.

Collins, R. L., Taylor, S. E., Wood, J. V., & Thompson, S. C. (1988). The vividness effect: Elusive or illusory? *Journal of Experimental Social Psychology, 24*(1), 1–18. https://doi.org/10.1016/0022-1031(88)90041-8

Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making, 21*(3), 241–251. https://doi.org/10.1002/bdm.586

Frederick, S., & Mochon, D. (2011). A scale distortion theory of anchoring. *Article in Journal of Experimental Psychology General.* https://doi.org/10.1037/a0024006

Furnham, A., Boo, H. C., & McClelland, A. (2012). Individual differences and the susceptibility to the influence of anchoring cues. *Journal of Individual Differences, 33*(2), 89–93.

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

Gilovich, T., & Griffin, D. (2002). Introduction—Heuristics and biases: Then and now. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 1–18). Cambridge University Press.

Grayson, C. E., & Schwarz, N. (1999). Beliefs influence information processing strategies: Declarative and experiential information in risk assessment. *Social Cognition, 17*(1), 1–18. https://doi.org/10.1521/soco.1999.17.1.1

Haddock, G., Rothman, A. J., Reber, R., & Schwarz, N. (1999). Forming judgments of attitude certainty, intensity, and importance: The role of subjective experiences. *Personality and Social Psychology Bulletin, 25*(7), 771–782. https://doi.org/10.1177/0146167299025007001

Hahn, U., & Harris, A. J. L. (2014). What does it mean to be biased. Motivated reasoning and rationality. In *Psychology of learning and motivation—Advances in research and theory* (Vol. 61, pp. 41–102). Academic Press Inc. https://doi.org/10.1016/B978-0-12-800283-4.00002-2

Harris, A. J. L., Blower, F. B. N., Rodgers, S. A., Lagator, S., Page, E., Burton, A., Urlichich, D., & Speekenbrink, M. (2019). Failures to replicate a key result of the selective accessibility theory of anchoring. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000644

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin, 135*(4), 555–588. https://doi.org/10.1037/a0015701

Heine, S. J., Proulx, T., & Vohs, K. D. (2006). The meaning maintenance model: On the coherence of social motivations. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc, 10*(2), 88–110. https://doi.org/10.1207/s15327957pspr1002_1

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). The Guilford Press.

Holman, B., & Bruner, J. (2017). Experimentation by industrial selection. *Philosophy of Science, 84*(5), 1008–1019. https://doi.org/10.1086/694037

Hooker, B., & Streumer, B. (2004). Procedural and substantive practical rationality. In A. R. Mele & P. Rawling (Eds.), *Oxford handbook of rationality* (pp. 57–74). Oxford University Press.

Jang, S. M. (2014). Seeking congruency or incongruency online? *Science Communication, 36*(2), 143–167. https://doi.org/10.1177/1075547013502733

Jung, M. H., Perfecto, H., & Nelson, L. D. (2016). Anchoring in payment: Evaluating a judgmental heuristic in field experimental settings. *Journal of Marketing Research, 53*(3), 354–368. https://doi.org/10.1509/jmr.14.0238

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated Closing of the Mind: "Seizing" and "Freezing." *Psychological Review*, *103*(2), 263–283. https://doi.org/10.1037/0033-295X.103.2.263

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480.

Larsen, K. (1990). The Asch Conformity Experiment: Replication and Transhistorical Comparisons. *Journal of Social Behavioral and Personality, 5*.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 551–578. https://doi.org/10.1037/0278-7393.4.6.551

Marcus, R., & Philipp, Y. H. (2017). The Resilient Personality Prototype Resilience as a Self-Deception Artifact? *Journal of Individual Differences, 38*(2), 1–11. https://doi.org/10.1027/1614-0001

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Mochon, D., & Frederick, S. (2013). Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes, 122*(1), 69–79. https://doi.org/10.1016/j.obhdp.2013.04.002

Molden, D., & Higgins, E. T. (2012). Motivated thinking. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 390–409). Oxford University Press.

Molden, D. C., & Maner, J. K. (2013). How and when exclusion motivates social reconnection. In *Oxford library of psychology. The Oxford handbook of social exclusion* (pp. 121–131). Oxford University Press.

Moore, T. J. (1995). *Deadly medicine: Why Tens of thousands of heart patients died in America's worst drug disaster*. Simon & Schuster.

Mussweiler, T., & Strack, F. (2000). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology, 78*(6), 1038–1052. https://doi.org/10.1037/0022-3514.78.6.1038

Neuberg, S. L., & Fiske, S. T. (1987). Motivational Influences on Impression Formation: Outcome Dependency, Accuracy-Driven Attention, and Individuating Processes. *Journal of Personality and Social Psychology, 53*(3), 431–444. https://doi.org/10.1037/0022-3514.53.3.431

O'Connor, C., & Weatherall, J. O. (2019). *The Misinformation Age: How False Beliefs Spread*. Yale University Press.

Open Science Collective. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Reber, R. (2017). Availability. In R. F. Pohl (Ed.), *Cognitive illlustions: Intriguing phenomena in thinking, judgment and memory* (2nd ed., pp. 185–203). Routledge, Taylor & Francis Group.

Reyes, R. M., Thompson, W. C., & Bower, G. H. (1980). Judgmental biases resulting from differing availabilities of arguments. *Journal of Personality and Social Psychology, 39*(1), 2–12. https://doi.org/10.1037/0022-3514.39.1.2

Rogers, P., Qualter, P., & Wood, D. (2016). The impact of event vividness, event severity, and prior paranormal belief on attributions towards a depicted remarkable coincidence experience: Two studies examining the misattribution hypothesis. *British Journal of Psychology, 107*(4), 710–751. https://doi.org/10.1111/bjop.12173

Rotliman, A. J., & Schwarz, N. (1998). Constructing Perceptions of Vulnerability: Personal Relevance and the Use of Experiential Information in Health Judgments. *Personality and Social Psychology Bulletin, 24*(10), 1053–1064. https://doi.org/10.1177/01461672982410003

Sanitioso, R., Kunda, Z., & Fong, G. T. (1990). Motivated Recruitment of Autobiographical Memories. *Journal of Personality and Social Psychology, 59*(2), 229–241. https://doi.org/10.1037/0022-3514.59.2.229

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112*(3), 610–628. https://doi.org/10.1037/0033-295X.112.3.610

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61*(2), 195–202. https://doi.org/10.1037/0022-3514.61.2.195

Shedler, J., & Manis, M. (1986). Can the availability heuristic explain vividness effects? *Journal of Personality and Social Psychology, 51*(1), 26–36. https://doi.org/10.1037/0022-3514.51.1.26

Slotter, E. B., & Gardner, W. L. (2009). Where do you end and i begin? Evidence for anticipatory, motivated self-other integration between relationship partners. *Journal of Personality and Social Psychology, 96*(6), 1137–1151. https://doi.org/10.1037/a0013882

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implication for understanding individual differences in reasoning. In *The Oxford handbook of thinking and reasoning* (pp. 433–455). Oxford University Press.

Stanovich, K. E. (2021). *The bias that divides us: The science and politics of myside thinking*. MIT Press.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking*. MIT Press.

Teigen, K. H. (2017). Judgments by representativeness. In R. F. Pohl (Ed.), *Cognitive illlustions: Intriguing phenomena in thinking, judgment and memory* (2nd ed., pp. 204–222). Routledge, Taylor & Francis Group.

Teovanović, P. (2019). Individual differences in anchoring effect: Evidence for the role of insufficient adjustment. *Europe's Journal of Psychology, 15*(1), 8–24. https://doi.org/10.5964/ejop.v15i1.1691

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability122. In *Cognlttive psychology* (Vol. 5).

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Tversky, A., & Kahneman, D. (2002). Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 19–48). Cambridge University Press.

Ušto, M., Drače, S., & Hadžiahmetović, N. (2019). Replication of the "Asch Effect" in Bosnia and Herzegovina: Evidence for the moderating role of group similarity in conformity. *Psychological Topics, 28*(3), 589–599.

Wright, W. F., & Anderson, U. (1989). Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. *Organizational Behavior and Human Decision Processes, 44*(1), 68–82. https://doi.org/10.1016/0749-5978(89)90035-6

Chapter 6
How We Were Made: The Evolutionary Origins of Thought

So we have examined human psychology from several angles: the accuracy of our beliefs, the accuracy of our metacognition, and the heuristics by which we form our beliefs. Each angle paints a picture of our truth-seeking abilities that is less than optimal—to say the least.

What might explain, then, why our psychologies are often sub-optimal at seeking the truth?

Psychologists have offered various explanations, with two competing explanations receiving significant attention in a recent influential book *The Enigma of Reason* (2017) by cognitive scientists Hugo Mercier and Dan Sperber. In this chapter, I will present these two explanations and Hugo and Mercier's critique of them, and then I will critically examine some aspects of their arguments.

6.1 Evolution, Functions, and the Intellectualist View

One of historically influential view of our reasoning faculties is what Mercier and Sperber call the "intellectualist tradition" (Mercier & Sperber, 2017). According to Mercier and Sperber's (2017) description, the intellectualist view maintains that "reason evolved to help individuals draw better inferences, acquire greater knowledge, and make better decisions" (p. 182). Put simply, then, the intellectualist view is that reason has the function of producing true judgments and making good decisions. Let us unpack what this means.

The word "function" has a very specific meaning in evolutionary theory. As evolutionary psychologist David Buss (2015) says, "The *function* of an adaptation refers to the adaptive problem it evolved to solve, that is, precisely *how* it contributes to survival or reproduction" (p. 36). For example, some butterflies have wings with patterns that disguise them from prey like birds; think of a butterfly with wings that look like the bark of a tree, for instance. (In biology, this is called *crypsis*, a

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0_6

technical term referring more or less to "camouflage".) In such a case, the function of the butterfly's wing pattern is to disguise it from prey. The reason that this is the pattern's function is because this disguise is how the pattern contributed to the butterfly's survival and reproduction.

How, then, do these functions arise? According to dominant evolutionary theory, they arise via random mutations. More specifically, when members of a species reproduce to create offspring, the fusion of the members' DNA will result in new mutations in the new offspring's DNA, mutations that are not present in the parents' DNA and which Buss calls a "copying error". The rest of the story is well told by Buss:

> Initially, a *mutation*, a copying error in a piece of DNA, occurs in a single individual. Although most mutations hinder survival or reproduction, some, by chance alone, end up helping the organism survive and reproduce. If the mutation is helpful enough to give the organism a reproductive advantage over other members of the population, it will be passed down to the next generation in greater numbers. In the next generation, therefore, more individuals possess the characteristic that was initially a mutation in a single person. Over many generations, if it continues to be successful, the mutation will spread to the entire population, so every member of the species will have it. (Buss, 2015, p. 37)

On this story, the butterfly's disguising pattern arose first as a result of random mutations. These mutations gave an evolutionary advantage to the disguised butterflies, because disguised butterflies are more likely to survive and reproduce than those butterflies who are visible to prey and are eaten. This then enabled the disguised butterflies to reproduce to make even greater numbers of disguised butterflies, eventually dominating the entire species of that particular kind of butterfly.

Every believer in evolution would then accept two things. The first is that random mutations led by a long process to the development of reason in humanity's ancestors. The second is that reason was so advantageous that it (eventually) dominated the species.

However, what differentiates psychologists is their views about *how* reason conduced to survival and reproduction—that is, on *what* the evolutionary function of reason is.

According to the intellectualist view, reason conduced to survival and reproduction by enabling us to get true beliefs and make better decisions. Perhaps, for example, an intellectualist might think our ancestors used reason to figure out answers to various questions: Which foods are poisonous or safe to eat? How can we find food or water? How can we treat and prevent illnesses? How can we safely raise children? The intellectualist might think that if reason helped us answer these questions correctly and act appropriately, we would be more likely to survive and reproduce than if we didn't answer them correctly—that is, than if we did not know which foods are poisonous to eat, how to find food or water, and so on.

Many psychologists and philosophers seem to think some kind of good reasoning conduced to survival and reproduction in this way. For instance, in agreeing with Daniel Dennett, Jerry Fodor (1981) states, "Darwinian selection guarantees that organisms either know the elements of logic or become posthumous" (p. 121).

6.2 Mercier and Sperber's Interactionist Approach

Mercier and Sperber have an influential discussion of the intellectualist view and their proposed replacement for it. In describing their views, I will quote copiously from their work, in part to minimize the possibility of false interpretations which inform my critique later.

They say that most philosophers and psychologists they have talked to endorse the standard "intellectualist" view of reason. But they think that it's the wrong view; it is "little more than hand waving" that "seems to point in a wrong direction" (Mercier & Sperber, 2017, p. 3).

They think this because the intellectualist view says reason serves a function that, in fact, it often systematically serves poorly: if the function of reason is to get true beliefs or make good decisions, why do pervasive "flaws" of reason like confirmation bias so severely compromise our truth-seeking and decision-making?

Their answer is that simply that the function of reason must be something else: "A biological mechanism described as an ill-adapted adaptation is more likely to be a misdescribed mechanism. Reason as standardly described is such a case" (Mercier & Sperber, 2017, p. 4).

Instead, they think reason must have a different function—or rather, two different functions.

The first function concerns justification; it concerns how we justify our beliefs and actions to others, and how we evaluate the justifications from others. This function, they claim, was selected by evolution because it solved an evolutionary problem of cooperation between our ancestors. More specifically, cooperation is essential to our species, but to cooperate, we need to determine who is reliable. When determining who is reliable, our direct observations of others is "limited to one's observations and is typically open to a variety of interpretations" (Mercier & Sperber, 2019, p. 72). For that reason, to determine reliability, we instead rely on the testimony of others, and this determines a person's reputation. A good reputation is then necessary for social success and biological fitness. Consequently, we have strong incentives to "protect and improve our reputation by explaining and justifying ourselves" (Mercier & Sperber, 2019, p. 72). We also have a strong incentive to "evaluate and possibly challenge these self-justifications" when they come from others (Mercier & Sperber, 2019, p. 72). For that reason, they say "[p]roducing justifications and evaluating them is, we claim, one of the two functions of reason" (Mercier & Sperber, 2019, p. 72).

The second function concerns argumentation; it concerns how we present arguments about what others should think or do, and how we evaluate the arguments from others. For Mercier and Sperber, argumentation is something that can occur successfully regardless of whether there is trust between the arguing parties. This helps us humans to engage in "much finer grained discrimination between valuable, and inaccurate and harmful messages," thereby enabling us to receive useful information and change our minds even when our "trust in the source would not have been strong enough for us to do so" (Mercier & Sperber, 2019, p. 72).

They call this the "interactionist" account—presumably because their posited functions have to more do with social interactions rather than with solitary intellect. How, then, do they support their account? They describe their general methodology for supporting their account, an account that is essentially the "adaptive hypothesis" that reason is an adaptation to produce and exchange justifications and arguments:

> There is much misunderstanding about the way to test adaptive hypotheses…What matters is the existence of a match between the function of an organ, or a cognitive mechanism, and its structure and effects. Do the features of the eye serve its function well? By and large, yes. Do eyes achieve their function well? By and large, yes.
>
> We can use the same logic to guide our examination of the data on human reason. Do the features of human reason serve best the functions posited by the intellectualist or the interactionist approach? … Does reason help him discard misguided beliefs and reach sounder conclusions? (Mercier & Sperber, 2017, p. 205)

Elsewhere, they reiterate their criticism of the intellectualist approach: "A genuine adaptation is adaptive; a genuine function functions" (Mercier & Sperber, 2017, p. 331).

Put simply, then, their approach utilizes the key premise that functions function well—or at least well enough to some degree which will be a subject of my critique later. More specifically, if a putative function does not serve its function well, then this would be evidence that this putative function is not the right function. So, if reason had the function of seeking truth and making good decisions, then it would serve this function well. But confirmation bias and other supposed "flaws" in reason would undermine this function, thereby casting doubt on whether this really is the function of reason.

But in contrast, they think there is good evidence for their account since their functions function well, at least in comparison. In particular, if their account is correct and reason had primarily social functions, then reason would have different features when it serves different roles—when it serves the role of producing reasons for oneself versus when it serves the role of evaluating reasons produced by others.

Regarding the first role: they claim that if their account of reason is correct, then the production of reasons would be biased and lazy. It would be biased in the sense that people would produce reasons mostly for their own side. They also claim reason would be lazy in the sense that people would not be very demanding of their own reasons: they may produce poor reasons for their actions or views without subjecting them to much scrutiny, for example.

They claim this role can explain several features of reason.

First, they claim their account explains "an otherwise puzzling feature of reason," namely, myside or confirmation bias. They claim this bias would be expected because it preserves reputation and better convinces others: "We're not going to appear more rational by providing reasons why what we did was stupid; we're not going to convince someone by giving them arguments for their point of view or against ours" (Mercier & Sperber, 2019, p. 73).

Second, it explains why "a solitary reasoner generally fails to correct their own mistaken intuitions: as reasons pile up to support these intuitions, the reasoner even risks becoming more confident or more polarized" (Mercier & Sperber, 2019, p. 74).

Third, it explains why "[w]hen a solitary reasoner doesn't have a strong intuition to begin with, reason pushes them towards the decision that is most easily justified—whether it is an otherwise good decision or not" (Mercier & Sperber, 2019, p. 74).

Of course, one might object that if reason served the social function of justifying oneself and arguing for one's own perspectives, then we would expect reason to "produce very strong reasons, to better defend ourselves and convince others" (Mercier & Sperber, 2019, p. 73).

But they anticipate this objection. They claim that the reason we do not find strong reasons to justify ourselves and convince others is that this is a "cognitively burdensome task," and obtaining strong reasons is better achieved by getting feedback from others. That way, we might lazily produce reasons that are "pretty banal" at first, although we can "adapt" from feedback and counter-arguments from others (Mercier & Sperber, 2019, p. 74).

They then argue that if their account is correct, reason would have different features when it serves the other role—the role of evaluating reasons produced by others. In particular, they expect reason to be unbiased and demanding in evaluating others' opinions rather than biased and lazy. Because it is unbiased, it accepts good reasons from others, even when those reasons "challenge our prior beliefs and come from sources we do not completely trust" (Mercier & Sperber, 2019, p. 74). Because it is demanding, it will reject weak reasons from others so as to not make us unjustifiably convinced.

They claim that this explains other kinds of evidence.

First, it explains why "the exchange of reasons allows good ideas to spread, and performance to increase, has been observed with a wide variety of contexts, from logical problems to forecasting or medical decisions" (Mercier & Sperber, 2019, p. 74).

Second, it explains why, when it comes to morality, "good reasons can change people's minds, even on moral and emotional matters" (Mercier & Sperber, 2019, p. 75).

Third, it explains the history of science which shows that scientists are those who are "in constant exchange (or at least revisiting past exchanges and anticipating new ones), who push each other to develop better arguments, and whose theories can rapidly take over a field, as soon as they are supported well enough" (Mercier & Sperber, 2019, p. 75).

Ultimately, then, they think the case for their account is demonstrably superior while the intellectualist approach flounders terribly.

As they say:

> Actually, the usual defenses of the intellectualist approach to reason are themselves good examples of biased and lazy reasoning. It is an undisputed fact that individual reasoning is rarely if ever objective and impartial as it should be if the intellectualist approach were right. In discussing what to do with this mismatch between theory and evidence, the possibility that the approach itself might be mistaken is rarely considered. Failures of reasoning are lazily explained by various interfering factors and by weaknesses of reason

> itself. Again, this doesn't make much evolutionary sense. A genuine adaptation is adaptive; a genuine function functions. (Mercier & Sperber, 2017, p. 330–1)

6.3 Critical Evaluation of Mercier and Sperber's Arguments

That is a summary of Mercier and Sperber's account, as well as the arguments they present for it. What, then, should we think of their account and their arguments?

Ultimately, I think their interactionist theory makes a valuable contribution to the literature. It is a possibly true account of the functions of reason, and indeed social forces may very well play an important part in evolutionary explanations of how reason came about. For that reason, I think their account, and its emphasis on social functions in particular, is an important viewpoint to consider.

That said, I am still open to both the intellectualist and interactionist approaches: I countenance both as live options but side firmly with neither.

But while I think Mercier and Sperber's theory may or may not be true, there are several issues in the arguments that they give for it.

The first is that their account is not obviously compatible with evolutionary theory, since evolutionary theory permits functional sub-optimality, but it is not obvious that their account does. According to evolutionary theory, biological organisms have functions which they sometimes carry out poorly—or at least not perfectly. A butterfly's wing pattern may have the function of disguising it from predators, yet it may successfully disguise it from some predators (such as particular birds), even if it fails to disguise it from others (such as particular lizards). Human teeth may have the function of enabling humans to chew food, even if some teeth (such as wisdom teeth) fail to help humans chew food and even if teeth cannot chew great quantities of every food (like sugary foods). Eyesight may have the function of helping humans see their environment, even if it fails to help them see their environment when it is too dark or too misty.

Time after time, organisms have functions that are imperfect, which work in some circumstances and may fail horribly in others.

As Buss (2015) notes, functions are not necessarily optimal:

> The concept of adaptation, the notion that mechanisms have evolved functions, has led to many outstanding discoveries over the past century…. This does not mean, however, that the current collection of adaptive mechanisms that make up humans is in any way "optimally designed." An engineer might cringe at some of the ways that our mechanisms are structured, which sometimes appear to be assembled with a piece here and a bit there. In fact, many factors cause the existing design of our adaptations to be far from optimal. (p. 17)

And as Gould (1997) states "even the strictest operation of pure Darwinism builds organisms full of nonadaptive parts and behaviors… All organisms evolve as complex and interconnected wholes, not as loose alliances of separate parts, each independently optimized by natural selection" (p. 52).

In determining whether something has a function, it is not so important whether it is optimal but rather whether it is *good enough* to have been selected for—good enough to have conferred *some* evolutionary advantage. Indeed, human eyesight is not optimal for seeing in general, but it is good enough to provide a strong evolutionary advantage compared to if we did not have it.

The question is, then, why cannot an intellectualist accommodate Mercier and Sperber's flaws by simply saying that reason—like many functions—is simply sub-optimal? Why couldn't reason be like a butterfly whose wings are good enough to disguise it from some predators, even if they are not good enough to disguise them perfectly from all predators? Why cannot reason simply be good enough at seeking the truth and making good decisions—good enough at conferring an evolutionary advantage—even if it sometimes succumbs to confirmation bias and other sup-optimal flaws?

It should be obvious that reason succeeds in finding the truth and making good decisions in many contexts, but it is also obvious that Mercier and Sperber take reason's failure in other contexts as evidence against its purported function of truth-seeking and decision-making.

What is not obvious, however, is how they can reconcile their attitude toward these failures and assert "a genuine function functions" with the fact that evolutionary theory entails that *many* functions actually function sub-optimally, if not poorly, in many contexts.

Perhaps Mercier and Sperber might respond that these other "sub-optimal" functions have been misdescribed because their description needs to be more sensitive to the contexts. Eyesight, they might claim, does not serve the function of seeing one's environment; instead, it serves the function of seeing one's environment *in contexts where there is sufficient light, it is not misty, and so forth*. Likewise, a butterfly's wing pattern does not serve the function of disguising it from prey; instead, it serves the function of disguising from prey *in contexts where the prey are birds* (and not lizards). These biological features serve the former functions poorly, but they serve the latter functions well, because they are suited to those contexts. Perhaps, Mercier and Sperber's work has the resources to furnish such a response, since, in defending their own posited functions, they say "[b]iological devices also have normal conditions: the conditions to which they are adapted" (Mercier & Sperber, 2017, p. 247). They might claim that reason serves its function poorly, while these other apparently sub-optimal features (butterfly patterns, wisdom teeth and eyesight) actually do serve their function well in the *right contexts*—or in the *normal conditions*.

But the intellectualist response should be obvious: they could claim reason also serves its function well, but *in the right contexts*. In particular, they may say that reason serves the function of seeking the truth and making good decisions *when there is no conflicting motivational bias, or when it is not an unnatural laboratory problem, and so on and so forth.*

Whatever the case, it is unclear how Mercier and Sperber can accommodate the undeniable fact of functional sub-optimality without giving intellectualists an escape route from their criticisms.

This issue focuses on the mere possibility of sub-optimality *in general*: it argues that regardless of what we think about reason, or its optimality, or its function—it is not clear how Mercier and Sperber can accommodate the possibility of sup-optimality *in general* with their "genuine function functions" criticism.

But beyond that, we might have a second worry: we might think that even if the intellectualist account is true, we would *expect* failures of the sort that Mercier and Sperber focus on. After all, if eyesight, butterfly wings, teeth, and many other things have functions that fail, and sometimes horribly, then couldn't we also expect reason to fail—sometimes horribly—at seeking the truth and making good decisions, as with confirmation bias and the like?

This, I would think, is the standard position of cognitive psychologists in the heuristics and biases research program, perhaps including many intellectualists. As Gilovich and Griffin (2002) point out, the heuristics and biases account of thinking "recognizes the constraints imposed by an organism's evolutionary history, constraints that yield noteworthy imperfections in function" (p. 10).

It is unclear, then, how it is charitable to claim that "the usual defenses of the intellectualist approach to reason are themselves good examples of biased and lazy reasoning" and that it is "an undisputed fact that individual reasoning is rarely if ever objective and impartial as it should be if the intellectualist approach were right" (Mercier and Sperber, 2017, p. 330–1).

A third worry is that it is not clear that Mercier and Sperber's arguments are coherent since it is not clear that their functions function sufficiently well—even by their own lights. In particular, they claim that if their account is correct and reason evolved to, among other things, evaluate the arguments and justifications of others, then "[w]e should be able to recognize good reasons, even if they challenge our prior beliefs and come from sources we do not completely trust" (Mercier & Sperber, 2019, p. 74). Again, elsewhere, they claim that the "interactionist approach to reason predicts that people should be good at evaluating others' reasons, rejecting weak ones and changing their mind when the reasons are good enough" (Mercier & Sperber, 2017, p. 273). And again, they claim people "have to be able to reject weak arguments while accepting strong enough ones, even if that means completely changing their minds" (Mercier & Sperber, 2017, p. 264).

But they give their own examples to show people often do not recognize good reasons, especially in cases where they challenge their prior beliefs. In their book, they discuss how the exchange of reasons can fail when used polemically rather than cooperatively, and so they discuss the example of Linus Pauling who, they claim, believed vitamin C could effectively treat cancer and failed to change his views. This was the case even when others at the prestigious Mayo Clinic conducted a "large-scale, tightly controlled trial" (p. 206) which showed vitamin C had no effect and which should have changed his mind. In this case, it is unclear how a "genuine function functions" (p. 331) to make us "able to recognize good reasons, even if they challenge our prior beliefs" when, as they say, "Pauling did not reason objectively" (p. 208) after he actually was presented with good reasons to challenge his prior beliefs.

Furthermore, Pauling's failure was not the only one they discuss; another example is Alphonse Bertillon, a French policeman who concluded that a man named Albert Dreyfus was a spy for the Germans. According to Mercier and Sperber

(2017), Bertillon failed to revoke his conclusion when presented with strong evidence which "should have immediately changed his mind" (p. 242).

What's more, they claim that Pauling and Bertillon are not exceptions or rare cases. Rather they refer to "The Bertillon in All of Us" (p. 241), and they make it clear Pauling is no unique case either:

> Pauling may have erred further than most respected scientists in his unorthodox beliefs, but his way of reasoning is hardly exceptional—as anyone who knows scientists can testify, they are not paragons of objectivity. (p. 207)

Ultimately, then, it is not clear how they can coherently claim that a genuine function functions and that reason serves a function of evaluating other's arguments, a function which, by their own lights, it often serves so poorly.

And even if they find a way to reconcile their supposed functions with such functional failures, it is not clear that the intellectualist could not also reconcile their views in a similar way.

But there is another worry about the coherence of their account: they claim that reason is biased and lazy while simultaneously serving to build reputation. Mercier and Sperber (2017) claim that, in reason's justificatory use, its "function is to manage your reputation" (p. 123) so that you will be seen as a reliable partner in cooperation. Yet, they also claim reason is biased and lazy in the sense that "people mostly produce reasons for their side" and "are not very exigent toward their own reasons" (p. 235). As a result, reason often leads us into error instead of producing balanced and strong reasons.

But who would you rather cooperate with: someone who you know produces biased and weak reasons which they are not very "exigent" toward, or someone who is fairer and produces genuinely stronger reasons for the opinions they hold? The answer, I think, would be obvious: after all, if Pauling and Bertillon were as biased and lazy as Mercier and Sperber would claim they are, would this improve their reputation and make them look like better partners in cooperation? Of course, other aspects of their career (such as their correct discoveries) may have enhanced their reputation, but I also think that any apparent bias and laziness on their part would have, if anything, done quite the opposite and somewhat undermined their reputation.

In fact, I suspect many of us know people who have a reputation for producing biased and weak reasons for their views, and I suspect this very fact makes us see them as less than ideal partners to cooperate with, at least in the circumstances where their reasons are so biased and weak.

But if this is the case, and if "a genuine function functions," then it is not clear how Mercier and Sperber can coherently claim that reason boosts our reputation on the one hand while it lazily produces biased and weak reasons for our beliefs and actions on the other.

Of course, they claim that, in dialogue, being lazy is a more cognitively efficient way of correcting or improving our beliefs, since we can refine them after receiving feedback from others. However, it is not clear this solves the problem, because again, think about who you would rather cooperate with: someone whose *initial* opinions are weak and biased or someone who is fairer and has genuinely stronger reasons for their initial opinions. Would you rather cooperate with someone who you give positive feedback to, because their initial reasons are so strong and fair, or someone who you give negative feedback to, because their initial reasons are so weak, biased, and lazy?

What's more, if Mercier and Sperber are correct in claiming that we should be good at evaluating others' arguments and justifications, then this feedback process involving weak, lazy, and biased reasons might serve to undermine reputations, since we would be good at detecting just how lazy and biased other people are. So while people may sometimes show bias to preserve reputation, it is not clear this a function of reason that succeeds in the ways that Mercier and Sperber think.

There is also another worry about the coherence of their argument: this time a tension between the functions of belief on the one hand and of reason on the other. As they make very clear, they think that the intellectualist approach is wrong: reason's function is not to seek truth or make good decisions. Oddly, however, beliefs are different: beliefs *do* function in a way that requires truth. In arguing that humans are not wishful thinkers, they state the following:

> Our beliefs are supposed to inform us about the world in order to guide our actions. When the world is not how we would want it to be, we had better be aware of the discrepancy so as to be able to do something about it. Thinking that things are the way one wishes they were just because one so wishes goes against the main function of belief. (Mercier & Sperber, 2017, p. 245)

So they say beliefs have the function of informing us about the world. But presumably beliefs can do this only if beliefs are true, since false beliefs about the world typically cannot inform us about how the world is or guide our actions. Yet, if that is the case, and if beliefs have the function of informing us about the world, then they must also have the function of being *true* beliefs about the world—at least in a large proportion of cases.

But this creates a puzzle: how can beliefs have the function of informing us about the world—or the function of being true—when arguably one of the main things that produces beliefs—namely reason—serves a very different function? If beliefs serve the function of being true, then one of the main things that produces those beliefs should serve the function of producing true beliefs too. Otherwise, either the beliefs are not fulfilling their function (since reason produces false beliefs), or reason is fulfilling the function of seeking the truth (since reason produces true beliefs). There would be no tension if reason served the function of producing true beliefs, but this is the very intellectualist position which Mercier and Sperber so vehemently deny. Again, it is not obvious how to reconcile their views.

6.4 Tangential Interlude: The Harm of Confirmation Bias

So I have critically discussed Mercier and Sperber's account by raising some worries.

Beyond that, though, I have another tangential worry about their arguments: if people accept their arguments, it could make society a worse place. People might interpret their arguments (rightly or wrongly) as a kind of vindication of what are usually seen as "epistemic vices": confirmation bias (which they call "myside bias"), lazy reasoning and arguing to preserve one's reputation instead of seeking the truth. As they say, for instance, "when defending a point of view, the myside bias is a good thing. *It is a feature, not a bug*" (Mercier & Sperber, 2017, p. 219). If that is the case, then people who accept their arguments may be more likely to behave in these traditionally epistemically vicious ways—all with a feeling of normalcy and entitlement.

The problem with this is that these biases can lead to bad outcomes. I suspect many of us know people who stubbornly defend their opinions from counterarguments, thereby creating conflict and sustaining beliefs which deter them and others from the truth. Furthermore, confirmation bias might have even more drastic, sometimes deadly consequences. As Lilienfeld et al. (2009) state:

> Arguably, the bias most pivotal to ideological extremism and inter- and intragroup conflict is confirmation bias, the tendency to seek out evidence consistent with one's views, and to ignore, dismiss, or selectively reinterpret evidence that contradicts them
>
> ... Like most large-scale cult movements…, virtually all violent regimes fan the flames of extreme confirmation bias in their citizens, especially their youth, by presenting them with only one point of view and assiduously insulating them from all others. Under Hitler, the Nazi government effectively hijacked the educational system by mandating a uniform curriculum emphasizing Aryan superiority, Jewish depravity, and the necessity of racial purity and by subjecting teachers to a month of systematic training in Nazi principles…. Educational materials that "contradict[ed] German feelings" (Noakes & Pridham, 1983, p. 437) were expunged, and teachers who deviated from the party line were fired. In contemporary Saudi Arabia, which spawned 15 of the 19 September 11, 2001, hijackers, curricula must adhere strictly to Wahhabi Islamist principles, and textbooks are screened carefully by the government for conformity to these principles. A number of widely used Saudi elementary school textbooks exhort their readers to spread Islam across the world through jihad, describe all non-Islamic religions as false, and inform students that Allah turned Jews and Christians into apes and pigs …. (p. 391)

Confirmation bias and Mercier and Hugo's other "features" of reason may then have bad consequences, both by deterring people from the truth and by creating conflict or other negative societal outcomes.

To be fair to Hugo and Mercier, though, the confirmation bias discussed by Lilienfeld et al. (2009) might not be the kind of confirmation bias that they endorse as a "feature". They might say confirmation bias is good only when it leads to the *initial* formation and presentation of reasons, but after these reasons have been presented in debates, this confirmation bias disappears, and people become naturally good at evaluating the arguments of others, even when these arguments challenge their own views. If Mercier and Sperber do think this, it would be good if they clarified this in their written work.

That said, even if they do not think this, others might incorrectly interpret their arguments as supporting the more full-fledged kind of confirmation bias, the type that persists after the initial formulation of reasons—as was supposedly present for the Pauling and Bertillon cases I discussed earlier.

In any case, there are other scholars who claim confirmation bias is sometimes a good thing under the right conditions. In particular, Gabriel and O'connor (2022) ran simulations of people interacting and updating their beliefs on new evidence. They found that groups with moderate levels of confirmation bias did better than particular other models without confirmation bias: in particular, these groups tended to reach "accurate consensus more often" (p. 10). However, they explicitly mention that the reason for this is different than the mechanism posited by Mercier and Sperber. According to Mercier and Sperber (2017), confirmation bias can be useful in a group setting because it leads to an "efficient way of dividing cognitive labor, with each individual finding arguments for her side and evaluating arguments for the other side" (p. 228). But while Gabriel and O'connor's (2022) results do not undermine this mechanism, it is not what their results support either; instead, moderate confirmation bias was beneficial because "confirmation bias leads to continued exploration and data gathering about multiple theories or actions" (p. 12). More specifically, they say "[d]ogmatic individuals force the group to more extensively test their options, and thus avoid pre-emptively settling on a poor one" (p. 2).

However, there are four problems with using this study to claim confirmation bias is a good thing simpliciter.

First, Gabriel and O'connor's (2022) study used a *very* specific definition of "moderate confirmation bias". In particular, in moderate confirmation basis, there is always a positive non-zero chance that an individual will accept some evidence that conflicts with their beliefs. However, whether they accept the evidence depends on two things. The first is how much they would have expected the evidence given their beliefs prior to encountering that evidence (sometimes called the *prior probability of the evidence*). The second is some tolerance factor t: a tolerance factor of 0 implies no confirmation bias, while higher values make one more likely to reject the evidence and not update their beliefs on it. For example, suppose someone thinks that a vaccine causes autism and that their tolerance factor t is such that $t = 2$. Now suppose their beliefs lead them to think it would be unlikely—with a probability of 5%—that a particular study of 1,000 children would fail to find higher rates of autism among those who took the vaccine. They then learn the results of the trial: it was not the case that there were higher rates of autism among those who took the vaccine. In this case, their probability of accepting that evidence is $0.05^2 = 0.0025$, so they are extremely likely to reject it and conclude it was, say, a poor and misleading study with no implications. In such a case, Gabriel and O'connor found such confirmation bias always helps the group reach accurate consensus in the long run (a long run which, as we will see, may be far too long a run). (They also found a similar result with a similar but more technical operationalization of confirmation bias.) Consequently, at best, their results suggest that this technical kind of confirmation bias benefits a group under specific conditions.

But this does not mean any kind of confirmation bias benefits a group, and that leads to the second problem: they found that "strong" levels of confirmation *harmed* the group. Again, strong confirmation bias is defined in a technical way: it occurs when the evidence is always rejected if it falls below a threshold of probability given one's belief. In this case, there is no tolerance factor t; instead, there is a parameter h such that all evidence that has a prior probability below h will be rejected. For example, suppose we take the vaccine example and set $h = 0.1$. With the moderate confirmation bias example, there was some probability of the vaccine skeptic accepting the evidence; it was just very, very low at 0.25%. But in the strong confirmation bias example, the vaccine sceptic would never accept the evidence, since the probability of the evidence given their prior beliefs was 5%, well below the *h* threshold of 10%. In such cases, polarization increases, the group is much less likely to reach consensus (whether it is accurate or not), and the bias leads "fewer individual actors, on average, to hold correct beliefs" (p. 16). So, strong confirmation bias can harm both a group and the individuals within it.

That said, they found even the moderate confirmation bias had its drawbacks, which leads me to the third problem. In particular, they found that "moderate confirmation bias always slows consensus formation, sometimes dramatically" (p. 17). The risk, then, is that moderate confirmation bias may be harmful when groups have sufficiently limited time to debate ideas and seek new evidence. And clearly, these limitations are often present; I suspect many of us know of cases where two or more parties disagreed, and they might have been able to reach consensus if they had enough time, patience and diligence, but, say, one or more of the parties just got too tired of arguing and gave up. Gabriel and O'connor's models would seem to imply that confirmation bias always prolongs debate and conflict. But, in real life, perhaps, it would be prolonged to a point where the parties give up and fail to resolve their disagreement before reaching any accurate consensus.

Given the infirmities of either kind of confirmation bias, then, is there an alternative that may speed consensus formation while preserving the benefits of moderate confirmation bias? I surmise that there is: fair open-mindedness—that is, when people accept the counterevidence to their opinions in a fair way but continue to open-mindedly explore the potential merits of hypotheses, even improbable ones, which the group would otherwise reject and fail to explore. After all, Gabriel and O'connor (2022) found moderate confirmation bias exerted its positive influence because it promoted "continued exploration and data gathering about multiple theories or actions" (p. 12). But surely, this same mechanism could result from open-mindedness, maybe even by definition. Yet, if that is the case, then the benefits of confirmation bias can be preserved without the negatives.

That, then, is the fourth problem with claiming there is evidence for the benefit of confirmation bias: ultimately, fair-minded open-mindedness may minimize conflict and lead to faster consensus, all while preserving the benefits of confirmation bias which arose through further data gathering and exploration of the possibilities.

Ultimately, then, I think there is currently no good reason to promote confirmation bias, or laziness, or any of the other traditional epistemic vices.

Instead, I think there are good reasons to avoid them: they can promote conflict and falsity, at least compared to fair open-mindedness, and they may even be responsible for societal problems such as prejudice, political polarization, and vaccine skepticism. And in chapter 7, we will see evidence that judgmental accuracy is increased by active open-mindedness, the opposite of confirmation bias (at least on some interpretations of it).

Of course, however, any potential societal harms of Mercier and Sperber's arguments are not necessarily reasons to reject those arguments, but at the very least, they should caution us about accepting them uncritically, and I have presented other independent reasons to worry about the strength of those arguments.

6.5 Summary

So we have considered two prominent evolutionary explanations of the infirmities of human thinking, particularly the inaccuracy and biases of our judgment.

Both explanations postulate that reason has specific functions—that is, that reason played various roles that were favored by natural selection because they conduced to the survival and reproduction of our ancestors.

According to the intellectualist explanation, reason has the function of seeking truth and making better decisions. This conduced to survival and reproduction by, say, helping our ancestors to determine which foods were safe to consume, how to locate water, how to treat and prevent illnesses, and so on. According to intellectualist view, reason is not perfect at making accurate judgments and good decisions, but it was good enough to confer an evolutionary advantage on our ancestors.

Mercier and Sperber think the intellectualist approach fails because it struggles to explain confirmation bias and other supposed flaws of human judgment and decision-making. After all, they claim that a "genuine function functions," and if reason appears to not function well enough, it is likely that its function has simply been misdescribed.

Instead, their interactionist explanation posits that reason serves the function of producing and exchanging arguments and justifications. For them, producing justifications helps us to explain our thoughts and behaviors to others in ways that preserve our reputation and make others see us as fit partners for cooperation. And they think arguments help us to persuade others about what to think or do. Reason is also competent, they claim, at evaluating justifications and arguments so we can discern who to cooperate with and which arguments to accept, even if the arguments come from sources we would not otherwise trust.

They cite various items of evidence in support of their account.

In particular, they claim that if the interactionist account is true, then we would expect reason to be biased and lazy when producing reasons. It would be biased toward producing reasons in our own favor because we will not enhance our reputation by producing reasons which undermine ourselves and our reputation. It would be lazy because this is a less cognitively burdensome behavior than producing strong arguments, especially when it is more efficient to have our views refined through dialogue and feedback from others. This expectation, they think, can explain why humans engage in myside or confirmation bias, why they fail to correct their own intuitions when alone and why reason pushes people toward the most easily justified decision when they initially lack strong intuitions.

They also claim that if their account is correct, then reason would be unbiased and demanding when evaluating reasons from others. This is so that reason can detect who to cooperate with and which arguments are reliable. They think this explains various facts, such as why good ideas spread, why performance increases when humans exchange ideas in teams, why good reasons can change people's minds on even emotional or moral topics and why scientific progress is made by scientists exchanging ideas.

However, while the interactionist perspective may or may not be true, there are various issues with Mercier and Sperber's arguments. First, it is not clear that their account is consistent with evolution: evolutionary theory permits that functions function imperfectly, like eyesight and (wisdom) teeth, but it is not clear that their interactionist account does. Second, given the prevalence of imperfectly functioning functions, we might even expect reason's functions to function imperfectly even if the intellectualist account is true, and if that is the case, the intellectualist may be able to somewhat explain their supposed flaws of reason, like confirmation bias. Third, it is not clear that Mercier and Sperber's arguments are coherent, since they give examples where the putative function of reason functions poorly: namely, individuals who, by their own lights, appear incapable of accepting good arguments from others—individuals like Pauling, Bertillon, scientists in general, and possibly even "All of Us". Fourth, another point of potential incoherence is that it is not clear how reason could serve the role of building reputation if, at the same time, it is lazy and biased in ways that would be detected by other competent evaluators and in ways that may consequently undermine one's reputation. Fifth, they effectively claim that beliefs serve the function of being true, but it is not clear how beliefs could serve this function: either reason does not serve the function of producing true beliefs (which would then undermine their claimed function of beliefs) or instead reason does serve the function of producing true belief (which would contradict their arguments). Finally, I mentioned that there are reasons to worry that Mercier and Sperber's account may result in societal harm if accepted, since people might (rightly or wrongly) see their account as a vindication of biases that potentially polarize people and incline them to violence, prejudice, and other negative social phenomena.

In any case, ultimately our mental faculties are like our physical faculties; just as we might have imperfect eyesight, wisdom teeth or other physical faculties, so too might we have sub-optimal human judgment. Evolution did not make us perfect in this respect. But just as we can improve our sub-optimal physical faculties—like our eyesight with glasses, or our physical health with medical science—so too can we improve our sub-optimal human judgment, and science can tell us how to do that. That, then, is the topic of the rest of this book.

References

Buss, D. (2015). *Evolutionary psychology: The new science of the mind* (5th ed.). Pearson.

Fodor, J. A. (1981). *Representations: Philosophical essays on the foundations of cognitive science* (1st ed.). MIT Press.

Gabriel, N., & O'connor, C. (2022). *Can confirmation bias improve group learning?* http://philsci-archive.pitt.edu/20528/1/Confirmation_Bias_Project%2C%20ArXiV.pdf

Gilovich, T., & Griffin, D. (2002). Introduction—Heuristics and biases: Then and now. In Thomas. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), Heuristics and biases: The psychology of intuitive judgment (pp. 1–18). Cambridge University Press.

Gould, S. J. (1997). Evolution: The pleasures of pluralism. *The New York Review of Books*.

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? *Perspectives on Psychological Science, 4*(4), 390–398.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Mercier, H., & Sperber, D. (2019). Précis of the enigma of reason. *Teorema: Revista Internacional de Filosofía, 38*(1), 69–76.

Noakes, J., & Pridham, G. (Eds.). (1983). Nazism: A history in documents and eyewitness accounts, 1919–1945. Exeter, United Kingdom: Department of History and Archaeology, University of Exeter.

Chapter 7
What Correlates with Accuracy: The Empirical Epistemology of Optimal Cognition

So, as we have seen, psychologists and others have studied how humans think for decades. And they have often unearthed some disturbing findings. The previous chapters are a testimony to this: while we often have accurate judgments as humans, we also often falsely diagnose people in ways that kill them, we sentence innocent people to death, and we have inaccurate judgments about many other topics—all without realizing it. However, as others have pointed out, research on judgment has focused primarily on *typical performance*—how humans ordinarily reason in various circumstances (Mellers et al., 2015a, b). In contrast, less attention has been paid to *optimal performance*—how well humans could perform.

7.1 Empirical Epistemology

That changed in the past decade. In 2011, an intelligence organization of the United States launched a research program. That organization is the Intelligence Advanced Research Projects Activity, also known as IARPA. The purpose of IARPA is to invest in "high-risk, high-payoff research programs to tackle some of the most difficult challenges of the agencies and disciplines in the Intelligence Community (IC)" (*About IARPA*, n.d.). The research programs are "high-risk" in the sense that they aim to tackle difficult problems where there is a high risk that any developed solutions may not be successful. The programs are also "high-payoff" in the sense that if any such solutions are successful, they promise to greatly benefit the intelligence community.

The Aggregate Contingent Estimation (ACE) was one such program. It aimed to "dramatically enhance the accuracy, precision, and timeliness of intelligence forecasts for a broad range of event types" (*Aggregative Contingent Estimation*, n.d.). To do this, one research team in the program, the *Good Judgment Project*, crowdsourced thousands of predictions about future geopolitical events, such as the

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0_7

outcomes of elections and international conflicts. They then tracked the accuracy of these predictions over time and found that a select group of people were exceedingly accurate. These people became known as *superforecasters*, and they are the subject of Philip Tetlock's best-selling book, *Superforecasting: The Art and Science of Prediction*.

Importantly, though, the program uncovered insights into what made some people more accurate than others.

The wealth of insight from these and other studies can be regarded as an emerging kind of field: *empirical epistemology*. Philosophy and, in particular, epistemology are often regarded as the fields of study that provide normative guidance about how we should think and reason (Broome, 2013; Chater & Oaksford, 2012). Empirical epistemology would then carry on this same aim of providing normative guidance. However, it differs from much of traditional epistemology in its methodology: it aims to provide its guidance, not solely on the basis of a priori argumentation, but also on the basis of *empirical scientific studies in real-world contexts*. Thus described, the ACE program may constitute the origins of empirical epistemology, although empirical epistemology arguably resembles earlier strands of work in judgment and decision-making, naturalized epistemology and even David Hume's empiricist approach. But even if we do not think the program is the origin of empirical epistemology, it is certainly an unprecedented source of insight for it.

The studies in this field could also be regarded as a kind of normative cognitive science. Like cognitive science, such studies scientifically investigate how humans actually think using empirical methods. But unlike much of traditional cognitive science, there is an explicit emphasis on carrying out these studies for the purpose of improving thinking—specifically by informing norms about how we should think.

We can then regard the literature following the ACE program as being an example of empirical epistemology and normative cognitive science: normative guidance is provided on the basis of scientific studies into what variables conduce to accuracy in real-world contexts. Since the program, such guidance has been integrated into training programs that have been administered to individuals in intelligence organizations and elsewhere.

7.2 The Domain Generality of Empirical Epistemology

However, empirical epistemology relies on an important presupposition: that there are *domain general methods for improving our accuracy.* By "domain general," I mean methods of improving accuracy which are useful in *many diverse* domains—domains as diverse as politics, medicine, law, everyday topics, and perhaps even more technical fields like physics. The idea, then, is that some of the methods for improving judgmental accuracy in some domains also improve accuracy in others: for example, using statistics may improve judgmental accuracy in a domain like geopolitical forecasting, but it can also do the same in a different domain like medicine. Because of this presupposition, this book treats particular insights about what improves accuracy in, say, geopolitical forecasting as potentially relevant to many other domains.

But an important caveat though: this presupposition does not imply that *every* method applies to *every* field. For example, it does not imply that using statistics increases judgmental accuracy when it comes to proving, say, theorems about mathematics or deductive logic, since these are domains where statistics are often simply irrelevant.

Of course, it is difficult, if not impossible, to delineate the fields where particular methods are useful from the fields where they are not, but I am not sure this delineation is necessary either, at least for the purposes of this book. Here, I think it suffices for the reader to consider these methods and to make their own judgment calls about whether they could be potentially applicable to the domains they are concerned with.

So that clarifies what the presupposition of domain generality is—or what it means.

But why should we think this presupposition is true?

Admittedly, I know of no empirical studies which directly test this presupposition, so the evidence for it is limited in this respect.

Despite this, however, I think there are good arguments for the presupposition.

For a start, it already appears to be true to some extent in different domains. For example, both textbooks in medicine (Stern et al., 2020) and training courses in geopolitical forecasting (Chang et al., 2017) instruct humans to avoid confirmation bias and to also seek evidence which disconfirms preferred hypotheses. What we have, then, is one method of improving accuracy—avoiding confirmation bias—which has been recommended in domains as dissimilar as medicine and geopolitical forecasting. Likewise, other similar recommendations have appeared in both fields, including recommendations to use teams to form judgments (Mellers et al., 2015a, b; National Academies of Sciences, 2015), to use base rate statistics to inform probability estimates (Chang et al., 2017; Stern et al., 2020; Tetlock & Gardner, 2015) and to track the accuracy of judgments or provide feedback about such accuracy to those making the judgments (Jaspan et al., 2022; Mandel & Irwin, 2021; National Academies of Sciences, 2015; Zwaan & Hautz, 2019). Of course, not all of these scholarly recommendations are based on studies directly showing their improvement of accuracy, but they are often based on some type of empirical investigation, such as an analysis of why judgments have gone wrong and how they could have been avoided (e.g., National Academies of Sciences, 2015).

Furthermore, it makes sense that these methods would conduce to accuracy in diverse domains when we consider the *rationales* for the methods—that is, when we consider *why* the methods are useful in those domains. For example, it is obvious that if one is antecedently biased toward a hypothesis and only seeks out evidence for it, then they are less likely to arrive at accurate judgments in cases where alternative hypotheses are true and where they fail to seek out the evidence for those alternatives. This is true regardless of whether the domain is medicine, politics, or many other empirical fields. A similar story can be told when we consider other methods, such as forming teams, using base rate statistics, and so forth.

Additionally, if the methods that improve judgmental accuracy in one domain also improve judgmental accuracy in others, then we would expect to see individuals who are accurate across many domains, simply because their truth-conducive methods in one domain would help them in another. To some extent, though, this is what we already see. Consider user 5265, the remarkably accurate individual described in Chap. 2 who had near perfect calibration. When we look at the questions they made forecasts about, we can see they concern diverse domains. These include Russian elections (e.g., "Who will be inaugurated as President of Russia in 2012?"), Guatemalan elections (e.g., "Who will win the September 2011 Guatemalan presidential election, or will a run-off be needed?"), military outcomes (e.g., "Will opposition forces in Syria seize control of the Syrian city of Aleppo by 30 April 2013?"), economics (e.g., "Will the London Gold Market Fixing price of gold exceed $1850 on 30 September 2011?"), law ("What will the outcome of Bo Xilai's trial be?"), religious topics (e.g., "Who will be the next Pope?"), epidemiology (e.g., "Will there be a significant outbreak of H5N1 in China in 2012?"), and possibly even more remote areas like physics (e.g., "Before 1 April 2013, will substantial evidence emerge that Iran has enriched any uranium above 27% purity?"). Time after time, we see user 5265 producing calibrated judgments across a variety of domains and questions like these. This is what we would expect if things that improved accuracy in one domain could also improve them in others.

So, to summarize, there are three reasons to think there are domain general methods for improving judgmental accuracy. First, some evidence suggests some of the same methods are already useful in areas as diverse as politics and medicine. Second, the rationales for the methods (like avoiding confirmation bias) would themselves lead us to expect their applicability in diverse domains. Third, there are indications that individuals who are accurate in some domains are also accurate in a variety of others, exactly what we would expect if accuracy could be improved in ways that are domain general.

In any case, to some extent, a presupposition of domain generality is already present in cognitive science and epistemology. A large amount of cognitive science looks at how humans actually do think, what makes them worse, and how to get better, as is the case with the heuristics and biases research program. And a large amount of epistemology looks at how humans should think and how to think better, as is the case with the literature on so-called "Bayesian epistemology". However, it makes sense to ask which domains these fields are about. I think it is fair to say that they are about thinking *in general*: they examine how we do and should think across a variety of diverse domains. And this also can be seen from the fields themselves. For example, ideas from Bayesian epistemology are applied in fields as diverse as astronomy and physics on the one hand (Dorling, 1979) and philosophy of religion on the other (Swinburne, 2004). Likewise, cognitive science's heuristics and biases research program has yielded insights across a staggering variety of disciplines and contexts (Gilovich et al., 2002). It would be natural for these disciplines to accept, if they have not already, that various methods could improve accuracy across a variety of domains.

However, one might object to the presupposition and claim that judgmental accuracy can be compromised by a range of domain specific features. For example, recall Gross's (2017) discussion of the false conviction rate in Chap. 3. There, he pointed out that a range of factors lead to false convictions in different kinds of criminal cases. Perjury is a leading contributor of false convictions in child sex abuse cases, misleading forensic evidence is a leading contributor in drug crimes, and mistaken eyewitness identifications are a leading contributor in crimes involving rape or robbery. If so many different factors give rise to false judgments in different contexts, how can one claim that there are domain general methods for improving accuracy? (Of course, Gross does not make this objection, at least not so explicitly, but one could hypothetically make it on the basis of his studies and others.)

How, then, could one respond to this objection?

I think that the objection is right on the mark about one thing: in any given case, there are plausibly context specific factors that lead to inaccurate judgments. Gross (2017) illustrates this well in his discussion of how different factors contribute to false convictions for different crimes.

But this does not imply there can be no useful domain general methods.

To consider this, think about the analogy of physical health. Different diseases involve (or are caused by) different factors. For example, cancer occurs when a body's cells grow uncontrollably, and diabetes occurs when a person's blood sugar is too high. Both are very different diseases with different etiologies (at least at some level of description). But it is also common knowledge that a healthy diet and exercise can reduce the risk of various kinds of cancer and diabetes, despite their different etiologies.

Judgmental accuracy, I think, is similar. To some extent, inaccurate judgments can have different causes, at least at some level of description. For example, false robbery convictions may be caused by misleading eyewitness testimony, while false drug crime convictions may be caused by drug misclassification. To some extent, then, these inaccurate judgments have different causes. But despite this, both kinds of inaccurate judgments may be prevented with the same methods, such as active open-minded thinking, cognitive reflection, and statistical thinking. For example, if someone utilizes these methods, they may be more likely to not take the evidence at face value (such as eyewitness or drug classification testimony), to think of alternative explanations (such as misidentifications and misclassification), and to be less confident in otherwise false convictions. These generally useful methods could reduce the probability of inaccurate judgments which would otherwise be caused by different specific factors.

For that reason, while I think this objection has a grain of truth, I do not think it rules out the possibility of domain general methods for improving judgmental accuracy.

Instead, I think the three previously mentioned reasons can give us some confidence that there are such methods, and I hope to list what I think some of these methods are in this chapter and in the next.

7.3 Insights from Empirical Epistemology

We will now examine insights that have emerged from empirical epistemology.

Note, however, that some of the literature is more focused on guidance for *collectives* and *crowdsourcing*. For example, Atanasov et al. (2017) studied different ways that collectives can produce predictions about the future: prediction markets and prediction polls. Prediction markets essentially involve individuals making financial bets on the outcome of future events. These bets then putatively represent the “wisdom of the crowd”—the collective estimate as to the probability of the event occurring. Predictions polls, in contrast, are not explicitly financial. Instead, they involve groups of participants providing direct probabilistic predictions about the probability of future events. Furthermore, the kind of prediction polling studied by Atanasov et al. (2017) also involved scoring participants’ predictions, as well as feedback for the participants as to how well they were doing. The results of their studies indicate that prediction polls outperform prediction markets, at least when the prediction polling places participants in teams and uses particular statistical aggregation methods to produce collective predictions.

That is an example of a study focused on *collective* cognition, but much of the literature contains guidance that is applicable for individuals too.

We will now explore this guidance in more detail.

Various studies over the past decade have explored what conduces to accuracy—typically measured with so-called *Brier scores* (which we discussed and critiqued in Chap. 2). We could say that the literature has identified four categories of variables which correlate with accuracy:

1. **Situational variables** that concern features of one’s situation.
2. **Motivational variables** that concern one’s motivations.
3. **Cognitive variables** that concern how one seeks out information and draws inferences from it.
4. **Metacognitive variables** that concern how one assesses their own cognition.

It is important to realize that no one variable is necessarily highly correlated with accuracy. This is because accuracy may require not just *one* feature, but a *set* of features which contribute to one’s accuracy.

We can think about this with the analogy of basketball. Consider what makes a basketball player good. There are plausibly many variables which determine this: whether they can dribble well, get rebounds, shoot three-pointers, slam dunk, and think strategically, for instance. However, no one of these variables explains even a lot of what it takes to be a great basketball player. Instead, different players may have different constellations of qualities which conduce to success: Stephen Curry is terrible at slam dunks but is great at three-pointers, while Shaquille O’Neal was terrible at even free-throws but excellent at slam dunks. Some other players are great at three-pointers—better than Shaquille O’Neal—but not as remarkable when all things are considered.

In a similar way, there may be no one quality that suffices to find the truth, and instead, a constellation of qualities may be required. Consequently, no one variable would necessarily be strongly correlated with accuracy, but many of them could play an important role together.

In any case, let us examine what variables have correlated with forecasting accuracy.

7.3.1 Situational Variables

Researchers from the Good Judgment Project discovered experimentally that a number of situational variables improve accuracy.

Training
One such variable is training. In particular, the Good Judgment studies found that forecasters were more accurate when they received specific training (Chang et al., 2017; Mellers et al., 2014, 2015a, b, 2019). Some of the training interventions became known as the "CHAMPS KNOW" guidelines. Each of the letters in "CHAMPS KNOW" refers to a specific component of the guidelines. For example, "C" stands for "Comparison Classes," the directive to use statistics for reference classes when estimating probabilities. Most of these components will be discussed below individually.

Teaming
Another situational variable is teaming. The Good Judgment Project found that individuals who were placed in teams were more accurate than those who were not (Mellers et al., 2014). Teams afforded participants the opportunity to share knowledge, to discuss rationales, and to motivate and engage each other. And teams comprised of the best individuals—superforecasters—significantly outperformed other teams.

Accountability
A third situational variable is accountability.

Chang et al. (2017) examined how accountability impacts forecasting accuracy.

To do this, they divided forecasters into four groups. One group—the control group—was not held accountable for their forecasts in any way. Another group had *outcome accountability*: that is, forecasters were told that they would be evaluated on the basis of the *outcomes* of their reasoning processes—namely, the scores of their forecasts about future events. Another group had *process accountability*: they were told they would be evaluated on the basis of the *processes* by which they arrived at their predictions, regardless of the outcome. Forecasters would be evaluated on, say, whether they used reference classes when making their estimates or other aspects of the CHAMPS KNOW guidelines. A fourth group had *hybrid accountability*—that is, a combination of both outcome and process accountability.

The results suggested that all forms of accountability resulted in roughly equal increases in forecasting accuracy—at least compared to the no accountability condition. The outcome accountable group was the most accurate (but not by a large amount compared to the other accountability conditions). Additionally, the process and hybrid accountable groups were better at persuading others of their reasoning. (Here, the persuasiveness of some reasoning was measured in terms of how likely others were to update their forecasts after reading the reasoning.)

The results suggest that outcome and process accountability can improve accuracy, but process accountability can also make one better at communicating persuasive reasoning to others.

7.3.2 Motivational Variables

A second category of variable concerns the motivations of individuals.

Mellers et al. (2015b) aimed to understand what makes superforecasters so super. One ingredient, they claimed, is that super forecasters had higher degrees of motivation and commitment. They specifically discuss how superforecasters are committed to "cultivating their skills" (p. 277). (Technically, the evidence they present is also compatible with the possibility that forecasters were more committed to finding out the truth than to developing their skill, but the development of skill is merely a means to this end.)

They cite three pieces of evidence in support of their claim that superforecasters are exceptionally motivated (2015b).

First, they were more likely than others to read more news from a news reader for forecasters: in years 2 and 3, for example, they clicked on an average of 255 stories compared to others who clicked closer to an average of 55 stories.

Second, they updated their forecasts more frequently than others: in the first year, for instance, they made an average of 2.77 forecasts per question, whereas others only made an average of 1.47 forecasts per question.

Third, superforecasters attempted more questions than others, even before they were designated as "superforecasters": in the first year of the forecasting tournament, for example, they attempted 25% more questions than others. In saying that, it is not clear how to reconcile this third piece of evidence with a study from Atanasov et al. (2020) which reported that attempting more questions did *not* conduce to accuracy. Perhaps, Atanasov et al. (2020) found that a significant effect was not perceptible when examining accuracy among *all* forecasters, even though Meller et al. (2015b) found a significant effect when comparing subsets of them, one of which includes the superforecasters.

Regardless, these three items of evidence, Mellers et al. claim, suggest superforecasters had greater motivation than others.

7.3.3 Cognitive Variables

A third category of variables concerns *cognitive factors*—that is, variables about how we think, seek information, and draw inferences from it. The evidence suggests more intelligent people are more likely to be accurate (Mellers at al., 2015a), but what specific cognitive factors make a difference?

Selective Effort

As mentioned earlier, Chang et al. (2017) report that a specific training program improved the accuracy of forecasters, training which contained the CHAMPS KNOW guidelines.

The "S" in CHAMPS KNOW refers to a principle known as *Select Appropriate Effort* (Chang et al., 2017). This involved instructing participants to engage in *cognitive triaging*—to allocate their "effort where it is likeliest to pay off" (Chang et al., 2017, p. 615). They do not elaborate on precisely what this means, but presumably it refers to an idea discussed by Tetlock and Gardner (2015): focus your effort on answering questions that are tractable instead of questions that are too difficult to provide informative answers about. Perhaps, for example, some questions are about matters that are too complicated or too far in the future to make reliable inferences about. The guideline may then be to avoid such questions, at least if Tetlock's (2015) principles of forecasting are anything to go by.

Technically, this guideline does not make one more accurate qua enhancing their epistemic abilities—that is, their abilities to reach the truth. At best, it merely means that they refrain from having opinions when their abilities are substantially limited. Consequently, the epistemic value of this variable is questionable, but it is included in this review here for the sake of completeness.

Other variables, however, may be more epistemically significant.

Active Open-Minded Thinking

A second cognitive variable is called *active open-minded thinking* (Baron, 1993). This refers to the extent to which an individual considers evidence against their favored opinions, spends enough time on a question before giving up, and takes into account the opinions of others when forming their conclusions.

In a series of studies, Haran et al. (2013) asked participants to answer various questions. These included questions about, for example, the outcome of football games and the frequency with which a particular object type (such as a particular colored ball) appeared in an image. They also measured the participants' levels of active open-minded thinking.

They report that individuals who were more actively open-minded were more accurate in their judgments and less overconfident.

Mediation analyses suggested that this was because actively open-minded individuals were more likely to search for more information to inform their opinions, and their opinions were more accurate as a result.

This is also consistent with the results of Mellers et al. (2014). They report that participants were more accurate when they received training in avoiding confirmation bias, and one aspect of active open-minded thinking is avoiding such bias. Of course, however, it is difficult to tell precisely how much of the increase in accuracy is attributable to this or to other features of the training. (And importantly, this is also a caveat that must be born in mind for other components of the training which lack independent evidence of their efficacy.)

In another paper, Mellers et al. (2015a) also report that more accurate forecasters scored higher on measures of active open-minded thinking.

The influence of active open-minded thinking is also consistent with the results of Atanasov et al. (2020). They found that when forecasting, some forecasters were more likely re-enter probability estimates that were the same as their previous estimates. They referred to such re-entering as *confirmation propensity*. They found that those with higher confirmation propensity were less accurate than others. One possible explanation of this finding is that more accurate forecasters are more open-minded: they are more likely to change their mind than others. Of course, it is also possible that those with greater confirmation propensity were simply lazier and less aware of new evidence, and this could also explain their lower accuracy and their reduced tendency to change their estimates. But the finding is consistent with both of these explanations regardless.

In any case, the available evidence either supports or is consistent with the possibility that active open-minded thinking improves accuracy.

Dialectical Complexity

Another variable that is closely related to active open-mindedness is dialectical complexity which Karvetski et al. (2022) describe as involving "grappling with the cognitive tensions between competing perspectives" (p. 3). They found that more accurate forecasters were more likely to express dialectical complexity and use words like "however," "yet," and "unless."

Cognitive Control

As measured with the so-called *Cognitive Reflection Test* (Baron et al., 2015; Frederick, 2005), Mellers et al. (2015a) found that more accurate forecasters were likely to have higher levels of cognitive control—that is, the ability to override intuitively appealing but incorrect responses, to avoid jumping to conclusions and the engage in more prolonged, careful consideration that lead them to the correct answer.

Search Processes

A fifth cognitive variable concerns search processes. In the CHAMPS KNOW training, individuals were taught the principle of *Hunt for Information*—alluded to with the "H." This "taught forecasters how to find information for forecasting" (Chang et al., 2017, p. 615), but precisely what this involves is unclear. Regardless, it is plausible that people's accuracy is influenced by the way that people search for information and their efficacy in doing so.

Reference Class Reasoning

An additional cognitive variable includes the use of *reference classes*. The CHAMPS KNOW training includes a component called the principle of *Comparison Classes*—denoted by the "C" (Chang et al., 2017). This instructs participants to estimate probabilities using base rate frequency statistics for appropriate reference classes. This is also referred to as taking the "outside view," the idea being that participants should estimate the probability of a specific events by considering things that are "outside" of that event, such as other similar events that have happened in the past. For example, if one wanted to predict whether, say, political stability would occur after a military coup, they should look to military coups from other countries, considering the so-called "base rate" or proportion of cases where such military coups resulted in stability compared to those where they did not. Training involving this principle has reportedly improved accuracy (Chang et al., 2017). Furthermore, Tetlock and Gardner (2015) claim that superforecasters are especially good at this kind of reasoning. Building on this, Karvetski et al. (2022) also examined the rationales of forecasters using natural language processing, and they found that the more accurate forecasters were more likely to use comparison or reference classes in their rationales.

A tangential comment: some philosophers would be especially pleased to hear this, since a number of them have espoused principles of reasoning that invoke reference classes (Carnap, 1951; Kyburg & Teng, 2001; Pollock, 1990; Reichenbach, 1949).

Updating Probability Estimates

Another variable concerns the frequency with which people update their probability estimates. This is encapsulated in a principle called *Adjusting*, denoted by the "A" in CHAMPS KNOW (Chang et al., 2017). This involves participants continuously reviewing their predictions, taking into account new evidence as time unfolds. (Philosophers would probably refer to this as continually updating beliefs in light of new evidence.) Again, the training containing this component was found to be efficacious in improving accuracy (Chang et al., 2017). Mellers et al. (2015a) also report that a strong predictor of the accuracy is the frequency with which participants update their beliefs.

Mathematical and Statistical Models

An additional variable concerns the use of models. In the CHAMPS KNOW training, participants were taught to, where possible, utilize formal models that depicted time-series or cross-sectional patterns (Chang et al., 2017). This is denoted by the "M" in CHAMPS KNOW.

Averaging Estimates

Another variable concerns how participants use information from multiple sources. In particular, the Good Judgment training instructed participants to average the estimates from polls, models, and expert opinions, where available (Mellers et al., 2019).

Scope Sensitivity

Another cognitive variable concerns *scope sensitivity*.

Scope sensitivity refers to how sensitive people are to changes in relevant quantities when making estimates. For example, in one study, 105 participants were asked about their willingness to pay for specific activities that would save the lives of migratory birds, such as geese (Schkade & Payne, 1994). The results found that participants were willing to pay the same amount regardless of whether they were told the activities could save 2,000 birds' lives or 200,000 birds' lives. In this case, the participants demonstrated scope insensitivity: changing the quantity of lives that could be saved would not change their estimate of their willingness to pay specific amounts for those activities.

Superforecasters are not like this. For example, participants gave probability estimates that the exchange rate for the Euro and US dollars would exceed a particular amount before December 31, 2014 (Mellers et al., 2015b). The amount varied between two groups, with each group containing a mixture of superforecasters and regular forecasters. For one group, the amount was 1.38, and for the other group, it was 1.40. The results found that, unlike regular forecasters, superforecasters had significantly different estimates depending on which value was presented. In other words, they were more sensitive to the specific values involved in their estimates.

Small Updates

As mentioned, Mellers et al. (2015a) also report that more accurate forecasters update their beliefs more regularly.

Another study from Atanasov et al. (2020) found that accurate forecasters update their beliefs not only more frequently but also in smaller increments. They also found that training results in updating with smaller increments.

The say their results are consistent with two potential explanations. First, the training encouraged use of base rates which might diminish the influence of new evidence about the specific event. Second, the training encouraged forecasters to combine information from multiple sources, such as averaging probability estimates. This possibly enhanced their ability to consider opposing ideas and may have made them less susceptible to strong pulls from single sources of evidence.

They also considered another explanation: that training encouraged forecasters to seek more information, and this could make them more attentive to evidence which subtly affected their estimates. However, this explanation was not supported by their analysis since, for instance, incremental updaters were not particularly active information consumers and their higher accuracy resulted primarily from their initial estimates that were more accurate, not from later adjustments that were more accurate.

Subject-Specific Knowledge

Another cognitive variable concerns subject-specific knowledge.

Chang et al. (2017) describe training which provided participants with geopolitical knowledge: how to analyze political actors' goals, capabilities, and constraints; the norms and protocols of important institutions which affect geopolitical events; the importance of "bottom-up sources of influence," such as populist movements and cultural conflicts; and the limits of one's predictive ability given our unescapable uncertainty about reality. As mentioned, this training improved accuracy, but again, it is difficult to know how much of the improvement is attributable to this feature or to others.

Mellers et al. (2015a, b) also report that more accurate forecasters scored higher on tests of subject-specific knowledge—in this case, political knowledge.

Avoiding Overconfidence, Imagining Possible Future, and Decision Trees
The Good Judgment training also encouraged participants to "imagine possible futures, use decision trees, and avoid judgmental biases such as overconfidence and base rate neglect" (Mellers et al., 2019, p. 213). However, not much more information is available aside from this, so it is difficult to elaborate on precisely what these features involve.

Word Count
Karvetski et al. (2022) found that more accurate forecasters were more likely to have lengthier rationales, as measured by word count. They found, however, that this was not merely accidental: more accurate forecasters tended to more thoroughly consider the forecasting questions and various competing perspectives (as with dialectical complexity above), thus leading to more accurate forecasts and to lengthier rationales to express their reasoning.

7.3.4 *Metacognitive Variables*

A final category of variables concerns our *metacognition*—how we think about our thinking and, more specifically, how we evaluate its accuracy.

Feedback
Some studies have found that feedback can improve accuracy and reduce miscalibration (Callender et al., 2016; Saenz et al., 2019).

For example, Moore et al. (2017) report further support for the value of feedback. They describe an experimental study in which forecasters participated in a one-hour training session. In it, participants provided answers to various questions, along with levels of confidence in their answers. They then received feedback about the accuracy of their estimates, and they were warned about the risk of overconfidence. These participants showed a slight improvement in particular measures of accuracy and a substantial improvement in calibration but mainly by reducing confidence. Again, though, it is difficult to know precisely how much of this is attributable just to the feedback or instead to other components of the training.

Postmortems
Another principle involves undertaking postmortems on prior predictions (Chang et al., 2017). Unlike feedback on the accuracy of one's past estimates, this variable concerns analyzing and understanding the outcomes of specific predictions—not just collections of them. Essentially, this refers to the process of reviewing past predictions, thereby identifying successes, mistakes and potential ways to improve in each case. Tetlock and Gardner (2015) also report that superforecasters are adept at this.

7.3.5 What Does Not Correlate with Accuracy

Above, we have explored a range of variables that correlate with accuracy, but empirical epistemology also has some negative lessons: lessons about what does *not* conduce to accuracy.

In particular, Tetlock's studies provide a word of caution: many societal indicators of accuracy may in fact be inaccurate. To refresh your memory, Tetlock (2005) studied the accuracy of political experts who made predictions about future events. All these events were political in nature, the subject matter they supposedly were experts in. How well, then, did education, type of profession, or particular other metrics correlate with accuracy? His answer was essentially "not very well". He states:

> It made virtually no difference whether participants had doctorates, whether they were economists, political scientists, journalists, or historians, whether they had policy experience or access to classified information, or whether they had logged many or few years of experience in their chosen line of work. (p. 68)

He also found fame correlated with accuracy but *inversely*: more famous people were more likely to be inaccurate.

Tetlock's studies then caution us about relying on these metrics of accuracy in general, but more on this will be said in the next chapter.

7.4 Summary

In this chapter, we considered the concept of empirical epistemology—the study of how humans should think based on empirical scientific studies about what actually conduces to accuracy. We also considered the origins of some empirical epistemology, origins which lie in work funded by the US intelligence community.

The presupposition of empirical epistemology is that there are domain general methods to improve accuracy—that is, methods that are that are useful in not just one domain, but in many diverse domains like medicine, law, political forecasting, and so forth.

I argued that this presupposition is plausible for three reasons. The first is that already the same recommendations can be found in different domains: for instance, both the domains of geopolitical forecasting and medicine have seen recommendations about the utility of teaming, tracking accuracy, avoiding confirmation bias, and using statistics. The second is that the rationales for the methods make sense of why they would be domain general: for instance, one can see how avoiding confirmation bias and fairly considering evidence against one's preferred views would lead to more accurate judgments in many domains, not just one. The third is that there is evidence of individuals who are accurate across many domains, something that is expectable if methods that improved accuracy in one domain could also improve accuracy in others. For these reasons, there are likely numerous domain general methods to improve accuracy.

The rest of this chapter then outlined variables which correlate with improved accuracy. These include situational variables, motivational variables, cognitive variables about how one seeks out information and draws inferences from it, and metacognitive variables which concern how one assesses their own cognition. It also discussed negative lessons from empirical epistemology: that is, insights about what does not conduce to accuracy. The result is a wealth of insight about variables that can (or cannot) predict or improve judgmental accuracy.

References

About IARPA. (n.d.). Retrieved July 2, 2020, from https://www.iarpa.gov/index.php/about-iarpa

Aggregative Contingent Estimation. (n.d.). Office of the Director of National Intelligence. Retrieved March 28, 2018, from https://www.iarpa.gov/index.php/research-programs/ace

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. Prediction Polls. *Management Science, 63*(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes, 160*, 19–35. https://doi.org/10.1016/j.obhdp.2020.02.001

Baron, J. (1993). Why teach thinking?-an essay. *Applied Psychology: An International Review, 42*(3), 191–237.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284.

Broome, J. (2013). *Rationality through reasoning*. Wiley Blackwell.

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6

Carnap, R. (1951). *Logical foundations of probability* (T. S. Gendler & J. Hawthorne, Eds.; Vol. 3). Routledge and K. Paul.

Chang, W., Atanasov, P., Patil, S., Mellers, B. A., & Tetlock, P. E. (2017). Accountability and adaptive performance under uncertainty: A long-term view. *Judgment and Decision Making, 12*, Issue 6.

Chater, N., & Oaksford, M. (2012). Normative systems: Logic, probability and rational choice. In *The Oxford handbook of thinking and reasoning* (pp. 11–21). Oxford University Press.

Dorling, J. (1979). Bayesian personalism, the methodology of scientific research programmes, and Duhem's problem. *Studies in History and Philosophy of Science Part A, 10*(3), 177–187. https://doi.org/10.1016/0039-3681(79)90006-2

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Gilovich, T., Griffin, D. W., & Kahneman, D. (2002). Heuristics and biases: The psychology of intuitive judgment (p. 857). Cambridge University Press. https://searchworks.stanford.edu/view/4815978.

Gross, S. R. (2017). What we think, what we know and what we think we know about false convictions. *Ohio State Journal of Criminal Law, 14*(2), 753–786.

Haran, U., Ritov, I., & Mellers, B. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, 188.

Jaspan, O., Wysocka, A., Sanchez, C., & Schweitzer, A. D. (2022). Improving the relationship between confidence and competence: Implications for diagnostic radiology training from the psychology and medical literature. *Academic Radiology, 29*(3), 428–438. https://doi.org/10.1016/j.acra.2020.12.006

Karvetski, C. W., Meinel, C., Maxwell, D. T., Lu, Y., Mellers, B. A., & Tetlock, P. E. (2022). What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? *International Journal of Forecasting, 38*(2), 688–704. https://doi.org/10.1016/j.ijforecast.2021.09.003

Kyburg, H., & Teng, C. M. (2001). *Uncertain inference*. Cambridge University Press.

Mandel, D. R., & Irwin, D. (2021). Tracking accuracy of strategic intelligence forecasts: Findings from a long-term Canadian study. *Futures & Foresight Science, 3*(3-4), e98.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015a). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1–14. https://doi.org/10.1037/xap0000040

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015b). Identifying and cultivating Superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science, 10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Mellers, B. A., Tetlock, P. E., Baker, J. D., Friedman, J. A., & Zeckhauser, R. (2019). Improving the accuracy of geopolitical risk assessments. In H. Kunreuther, R. J. Meyer, & E. O. Michel-Kerjan (Eds.), *The future of risk management* (pp. 209–226). University of Pennsylvania Press.

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science, 63*(11), 3552. https://doi.org/10.1287/mnsc.2016.2525

National Academies of Sciences, E. (2015). *Improving diagnosis in health care*. https://doi.org/10.17226/21794.

Pollock, J. (1990). *Nomic probability and the foundations of induction*. Oxford University Press.

Reichenbach, H. (1949). *The theory of probability: An inquiry into the logical and mathematical foundations of the calculus of probability* (2nd ed.). University of California Press.

Saenz, G. D., Geraci, L., & Tirso, R. (2019). Improving metacognition: A comparison of interventions. *Applied Cognitive Psychology, 33*(5), 918–929. https://doi.org/10.1002/acp.3556

Schkade, D. A., & Payne, J. W. (1994). How people respond to contingent valuation questions: A verbal protocol analysis of willingness to pay for an environmental regulation. *Journal of Environmental Economics and Management, 26*(1), 88–109. https://doi.org/10.1006/jeem.1994.1006

Stern, S. D. C., Cifu, A. S., & Altkorn, D. (2020). *Symptom to diagnosis: An evidence-based guide* (4th ed.). McGraw-Hill Medical.

Swinburne, R. (2004). *The existence of god* (2nd ed.). Clarendon Press.

Tetlock, P. (2005). Expert political judgment: How good is it? How can we know? Princeton University Press.

Tetlock, P., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Broadway Books. https://doi.org/10.1201/b15410-25

Zwaan, L., & Hautz, W. E. (2019). Bridging the gap between uncertainty, confidence and diagnostic accuracy: Calibration is key. *BMJ Quality & Safety, 28*(5), 352–355. https://doi.org/10.1136/bmjqs-2018-009078

Chapter 8
How Can We Get More Accurate: Recommendations About Human Judgment

This book was a tour through a variety of literatures that are inspiring at times and then downright depressing at others. What they suggest on the one hand is that human judgment is often inaccurate, much more than many would hope or expect, and with severe consequences. But on the other hand, some of the literature yields insights about how we can improve the accuracy of our judgments.

My task in this final chapter is to try to synthesize these insights into practical recommendations.

The recommendations come in the following three categories:

1. Recommendations for improving our judgment as individuals.
2. Recommendations for estimating the accuracy of others.
3. Recommendations for conducting our organizations in ways that conduce to success.

That said, no single one of these recommendations *guarantees* success. As mentioned in the previous chapter, truth-seeking is arguably somewhat like playing basketball: it requires not one quality, but rather a constellation of qualities that conduce to success. These recommendations, then, are best interpreted as a *set* of ways by which to *raise the probability* of accurate judgments. Each recommendation may also require further thinking about how to implement it in the intended contexts.

So, without further ado, let us now go category by category and outline some recommendations, starting with the first category on improving our own judgments.

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0_8

8.1 Category 1: Improving Our Own Judgments

8.1.1 Foster Motivation

Successful truth-seekers are often motivated to find the truth. That much is suggested by the studies of forecasting which found that superforecasters—the super truth-seekers about political topics—looked at more evidence than others, updated their opinions more often, and attempted more forecasting challenges than others. All of these things imply exceptional degrees of motivation, especially since Good Judgment did not pay them for their participation.

So, if we want to have accurate judgments, we should foster the motivation to do so (if we do not have it already, that is).

Of course, to my knowledge, empirical epistemology has not furnished empirical data about precisely what can help us foster motivation, but some common-sense techniques may work. These include, for example, reminding yourself about why judgmental accuracy is important to you. Do you or others depend on having accurate judgments in some important way? We can think about the negative consequences of inaccuracy, for instance. If you are a doctor, could judgmental inaccuracy harm your patients or your career? If you are a juror, could judgmental inaccuracy cause you to inflict punishment on undeserving innocents, all while the true criminals go unpunished? We could also think about the positive consequences of accuracy. If you are a political expert, could judgmental accuracy help your country or organizations flourish in ways that would motivate you to reach more accurate judgments? Periodic reflection on questions like these may serve to foster and maintain the motivation to be accurate.

8.1.2 Become Accountable

Motivation for accuracy can also come from accountability.

As we saw, experimental studies suggest people are more likely to be accurate when the accuracy of their judgment is accountable to others. As a result, you may wish to explore ways in which you can become accountable to others.

Of course, this might sound daunting, especially if you are not so confident that you are very accurate. One solution to this may help though: what I call "practice periods".

Let me give an example of this. I recently taught an introductory philosophy of science course at Stanford. Western analytic philosophy, the tradition I have worked in, is often obsessed with conceptual clarity and argumentative rigor: having studied in four different disciplines in various universities, I would confidently say analytic philosophers focus on clarifying the meaning of terms and evaluating the rigor of arguments far more than any other discipline I am aware of (perhaps excluding mathematics). As a result, students who are new to philosophy often find it intimidating to step into an introductory class where the clarity and rigor of their work is scrutinized so carefully. And rightly so: as my students themselves will tell you, there is a lot that they "can't get away with" in philosophy courses compared to their other disciplines, and this requires some adjusting.

To alleviate intimidation, the students get "practice periods". That is, some of the assessments are ones where they will pass merely for attempting the assignment. They do not get a letter grade (like an A or B-) placed on their official academic record, but they receive feedback about what grade they *would* have gotten *if* their assignment received a grade on their record. This means that they have the opportunity to practice, to receive feedback and to try to improve on a new kind of assessment rubric without the fear of failure.

But the practice period is just a period; after some amount of practice, they will then be accountable, assessed, and formally graded on a letter basis. This increases their motivation to use the practice period as stepping-stones for improving until they will be eventually accountable.

The end-of-quarter survey feedback and informal class discussions suggested this helped students.

Of course, this is just anecdotal evidence, but my suspicion is that practice periods could work for similarly ensuring accountability in other contexts.

But while the evidence for practice periods is anecdotal, the evidence for the utility of being accountable is not: both common-sense and experimental studies suggest that accountability increases accuracy, as discussed in the previous chapter.

8.1.3 Track Your Accuracy

Being accountable, either to oneself or to others, requires some way of measuring accuracy—some way of keeping a track record of one's accuracy.

Tracking accuracy not only is necessary for accountability, but it provides feedback about one's performance, which some studies suggest can improve accuracy and reduce miscalibration, at least in some contexts (Callender et al., 2016; Moore et al., 2017; Saenz et al., 2019).

However, not all ways of tracking and measuring an accuracy are equally good. In Chap. 2, I reviewed various measures of accuracy, including widely used measures like so-called Brier scores and unbinned calibration. While these measures might be useful in some contexts, they have limitations: for example, unbinned calibration can make individuals appear more accurate than they really are, while Brier scores sometimes fail to distinguish good calibration with poor resolution from bad calibration with more resolution.

As mentioned in Chap. 2, my preferred ideal measures are binned calibration (and resolution) for measuring the accuracy of individuals and weighted binned accuracy (and resolution) for measuring the accuracy of groups. Of course, these sometimes require many probability judgments, so these may not always be feasible, and other measures may be preferable, albeit sub-optimal, in particular contexts.

That said, when it comes to less-than-certain judgments, we usually need to assess many such judgments to get a clearer picture of how accurate we are. As mentioned, even a perfectly accurate individual will be 90% confident of things that will be false 10% of the time. A small sample of such judgments may give an unrepresentative impression of someone's accuracy, especially if they happened to get lucky or unlucky.

The appropriateness of a measure may to some extent vary from context to context, and I explore relevant subtleties more in my longer answer to the question of how to measure accuracy (Wilcox, Working Paper A). Updates about measuring accuracy may be posted at my website too (as mentioned in the next chapter).

Regardless, if possible, we should measure accuracy. But measuring accuracy comes with another caution in the next recommendation.

8.1.4 Be Your Own Skeptic: Expect Inaccuracy and Embrace Humility

Once measurement produces your track record, what might you be able to expect?

Well, the available evidence suggests humans are less than ideally accurate in many contexts, but—unfortunately—that includes you too. But do not take this personally; it includes all of us, me as well. And time after time, juries convict innocent people, doctors misdiagnose patients, and political experts are certain of things that are false. All the available evidence suggests we are frequently inaccurate and, what's more, we do not even realize it.

This warrants a kind of cognitive humility: we all, yourself included, are likely going to get things wrong, possibly even things we are very confident about—maybe even certain! This, of course, does not mean that we cannot have some degree of confidence in our ideas. When we are confident of something, we may be right most of the time, as the evidence indicates is generally the case. And there are some domains where we are highly accurate. But the evidence also means we should expect and embrace the possibility we are wrong about a considerable fraction of the things we are confident about.

This may feel bad, but I think we need to get comfortable with the fallibility of human nature: humans are imperfect, especially in our judgment, and that is okay, provided we act appropriately.

Yet acting appropriately, I think, includes making an effort to improve the accuracy of our judgment.

And this can be done: the evidence has proved that some people are more accurate than others—some are even perfectly calibrated—and that accuracy can be substantially improved, at least in some contexts.

The rest of the suggestions in this category explore how this improvement can occur.

8.1.5 Beware of Intuition

The Good Judgment Project found that more accurate forecasters were more likely to avoid jumping to intuitive but false conclusions, even if they seemed obvious, and to instead engage in more prolonged and careful thinking. For example, consider this problem which you might have encountered before: a baseball bat and a ball cost $1.10 together and the bat costs $1 more than the ball. How much does the ball cost? The majority of people jump to the intuitive conclusion that the ball obviously costs 10 cents. But this is not correct. If the ball costs 10 cents, and the bat costs a dollar more at $1.10, then they would cost $1.20 together, but they do not. The correct answer is that the ball costs 5 cents, not 10 cents. More accurate forecasters are those people who are more likely to pause, to think more carefully about a situation, and to avoid intuitions that lead them seemingly obvious but actually incorrect conclusions like these. Put simply, they are more likely to override Type 1 thinking with Type 2 thinking, as per the discussion in Chap. 5.

The studies supporting this were correlational, but these correlations are unlikely to be spurious: experimental studies have repeatedly shown that intuition can lead us astray, and more prolonged, careful thinking can help us avoid jumping to falsehoods when they intuitively appeal to us.

8.1.6 Practice Active Open-Minded Thinking

That fallibility of our judgment—the inaccuracy of our judgment—should also make us open-minded, especially to the possibility of being wrong and to other ideas we might otherwise be confident are false.

But there is another reason to be open-minded: the evidence suggests “active open-mindedness” can improve our accuracy (Atanasov et al., 2020; Haran et al., 2013; Mellers et al., 2015a). Active open-mindedness here refers to searching for evidence against your favored opinions, taking into account alternative hypotheses and considering the perspectives of others.

Active open-mindedness, then, refers to a constellation of qualities that are the anti-thesis of confirmation bias, jumping to conclusions, and particular other putative errors in judgment.

So, if you want to be more accurate about a subject, try taking your preferred hypothesis and looking for evidence against it. Try to consider alternative hypotheses which might explain the evidence or data just as well. Try seeking out the perspectives of others, especially if they differ from yours.

All of this leads to another recommendation though.

8.1.7 Gather Subject-Specific Knowledge and from Diverse Sources

Many people have many opinions about many things, but not all of them are ones they will necessarily gather subject-specific knowledge about. And by gathering knowledge, I mean through, for example, newspapers, journals, books, videos, or other sources. People often have strong opinions from hearsay or a couple of personal experiences, but good truth-seekers are not always like this. It should be no surprise, then, that studies show good truth-seekers seek knowledge to inform their opinions (Mellers et al., 2015a, b), but they also do not settle for one source either. They seek information from multiple sources to triangulate and further inform their perspectives. For example, the superforecasters in years 2 and 3 of the ACE program clicked on an average of 255 stories compared to others who clicked closer to 55 stories on average (Mellers et al., 2015a, b).

So truth-seekers should seek knowledge, but what kind of knowledge?

8.1.8 Use Statistics, Especially Base Rates

In some training programs that successfully improved accuracy, participants were instructed to look for statistics (Chang et al., 2017). For example, if someone wanted to predict an election outcome, they should look at statistics about previous elections. Or if they wanted to predict whether, say, political stability would occur after a military coup, they should look to military coups from other countries, considering the so-called “base rate” or proportion of cases where such military coups resulted in stability compared to those where they did not. Such base rates can inform one’s initial estimate of the probability of a proposition. Then, one can look to other evidence more specific to the problem at hand in order to update their judgments. Considering base rates like these is also called taking the “outside view” of a question, and it is something that Tetlock and Gardner (2015) say superforecasters are especially good at. This was also confirmed when a recent study found that more accurate forecasters cited base rate statistics more frequently in the rationales for their forecasts compared to less accurate forecasters (Karvetski et al., 2022). Informative statistics are antidotes to ill-informed heuristics, like availability; where possible, we should use them.

8.1.9 Average Estimates from Conflicting Sources

However, sometimes, statistics, models, experts, and other sources of insight can conflict. For example, a pretty good opinion poll may suggest that a candidate has a 45% chance of winning an election, while another equally good one might give the candidate a 60% chance of winning. What should one do with such information?

Some training programs tell participants to average estimates where possible (Mellers et al., 2019), and such programs reportedly improve accuracy. Of course, it is not clear whether there is evidence to suggest that specifically *this* component of the training improves accuracy. Consequently, this recommendation is tentatively made with some caution. That said, it is consistent with the evidence, and it also seems sensible as a tentative recommendation on its own grounds.

8.1.10 Test for Scope Sensitivity

Having gathered our information and made judgments, we can test the rationality of our inferences in another way: by asking whether they are *scope sensitive*. As mentioned in Chap. 7, "scope sensitivity" refers to when our judgments are sensitive to things which should change our judgments if those things changed. A scope insensitive judger is someone who would, say, give the same amount of money to a charity which saves 2,000 birds' lives as they would to a charity which saves 200,000 birds' lives. In that case, their judgment about how much to give would be insensitive to changes in the number of lives saved, even though the quantity of lives saved merits different levels of contribution: a change in the number of lives saved should change how much we give to charity. On this note, superforecasters are more likely than others to be sensitive to scope changes (Mellers et al., 2015b). Their predictions about the projected exchange rate for Euros and US dollars would be sensitive to differences by just a cent or two, for example, whereas other less accurate predictions would be less sensitive. For that reason, if you make a judgment about a proposition, ask what would happen if your evidence changed or what would happen if the proposition changed: would your judgment be sensitive to the relevant changes, or would your judgment be insensitive to keys features of the situation which should affect your judgments?

8.1.11 Do Postmortems

So suppose you have followed the above steps, made some inferences, and tracked your accuracy.

After this, one can also improve their accuracy by doing "postmortems," so to speak—that is, by analyzing their past judgments and why they worked well or poorly. This can help identify trends to inform how to correct our judgment. Superforecasters are reportedly adept at this (Tetlock & Gardner, 2015), and accuracy has also been improved by training which instructs participants to do postmortems (Chang et al., 2017).

That said, we should also try to avoid overdiagnosing the cause of past failures: even a perfectly accurate individual will be 90% confident in things that are false

10% of the time. So, generally speaking, take one-off failures with a grain of salt and try to interpret outcomes in the broader context of a track record.

8.1.12 Take Some Training

Finally, as has been repeatedly mentioned, training in good thinking can improve thinking. Consequently, you may wish to enroll in courses or short workshops dedicated to improving accuracy. But a word of caution: many courses offer advice about how to improve, yet not all of the advice is equally good, and some courses are probably not worth your time. To distinguish good opportunities for training from others, look at how the training was devised: what studies informed the training, and is there evidence that the training actually works? And watch out for misleading advertising: companies frequently market questionable products based on "evidence" which either weakly supports a strong effect or strongly supports a weak to negligible effect. Updates about training programs may be posted at my website too (as mentioned in the next chapter)

8.2 Category 2: Estimating the Accuracy of Other Sources

Of course, no human is an island: we form judgments about the world, not just by relying on our own reasoning faculties but also by relying on the reasoning faculties of others. Many of the opinions we have about the world are mediated by the testimony from others; this includes beliefs about election results, diagnoses, treatments of diseases, and the entirety of science that is reported via journal articles. Society is necessarily like this: our individual and collective lives function much better when we can rely on accurate information from others.

But this creates a problem: how do we know what sources of information to trust?

In this section, I list some recommendations for how to estimate the accuracy of other sources.

8.2.1 Be Skeptical of Judgment, But Not Too Skeptical

Throughout this book, it was obvious that humans are frequently much more inaccurate than they or others would hope or expect.

For this reason, it pays to be skeptical: we should not automatically assume that others are accurate, or that they even realize whether they are accurate or not. As we saw in Chaps. 2 and 3, many people are even certain of things that are false, including experts whose job is to give opinions about political matters. And as we saw in Chaps. 3 and 4, often people are repeatedly incapable of estimating their own accuracy. For that reason, a dose of skepticism is warranted (although this admittedly might not help you win friends if you express it so explicitly).

Yet we should not be too skeptical. Even the people who were certain of things that were sometimes false were still certain of things that were mostly true. Some of Tetlock's political experts were certain about things that were false 19% of the time—which is pretty bad—but the other 81% of things they were certain about were true—which is still somewhat good. So humans are accurate in some contexts and to some degree; the problem is simply that we are often far from perfect in various contexts—and with severe consequences.

8.2.2 Estimate Accuracy Based on Track Records

How, then, can we estimate accuracy? Well, there are different ways, each with their strengths and limitations.

Yet, one way is salient: use track records, at least where it is possible. In another paper, I argue extensively for "calibrationism", the idea that judgments of probability are trustworthy largely to the extent that they are supported by track records of accuracy (Wilcox, work in progress). Many of the pioneering insights of empirical epistemology emerged from tracking accuracy and seeing what improves it. The evidence shows that *accurate* individuals often continue to be accurate, and the same is true for *inaccurate* individuals too (Mellers et al., 2015a, b), although some evidence also shows inaccurate individuals can also improve with practice (Wilcox, Working paper B).

So we know that, in fact, it is possible to both measure and predict accuracy from track records.

But these are not just any track records: not every way of measuring accuracy works just as well.

Yet some ways do work, and they include track records of *calibration* and *resolution,* as discussed in Chap. 2. So, if a doctor makes confident diagnoses, how frequently were those correct? If someone makes confident predictions, how frequently did those predictions come true? In measuring calibration, we need to ask whether such levels of confidence vary appropriately with such frequencies of truth. In measuring resolution, we need to ask how informative those predictions were in the sense of being close to confidence levels of 0% or 100%.

Of course, when estimating someone's accuracy, you might be thinking, "But there are no track records! We don't have enough information about someone's past accuracy!"

And if so, you are right: there often are no track records, and that is a big problem in society today. This is why I argued in Chap. 3 that much of society simply flies blind; we simply trust others without having any good evidence about how accurate they actually are. And in the contexts where we have tracked accuracy to some degree, such as in medicine, political judgment, or law, we have grounds to be concerned not only about such contexts but also about these other contexts where we have not tracked accuracy so carefully.

But even if we lack robust track records of calibration, sometimes we can acquire useful information in other ways. Ideally, we would have individuals take calibration tests in the domains they make judgments about: that is, we would have them make a large number of probability estimates about those domains and then measure whether their confidence levels vary appropriately with the frequency with which things are true. But sometimes we can gather useful information about past success, even without such tests. For example, if a scholar cites articles, we can have a look at the literature and see how accurate their depictions are; that can indicate how accurate they are, at least in their depictions of the literature. If you see a doctor, see if you can find reviews online about how good or bad they are. Of course, all these sources of information are fallible and must be interpreted cautiously: false reviews can be written for medical clinics, for instance. But at the very least, such information can be at least something in the absence of highly informative track records.

Yet such track records are not only useful for estimating the accuracy of humans.

8.2.3 Look for Models or Theories with Track Records of Accuracy

Track records are also useful for estimating the accuracy of theories, models, or other kinds of computational techniques.

This was illustrated by Youyang Gu, a figure who rose to prominence during the COVID-19 pandemic. At the beginning of the pandemic, Gu was then a 26-year-old MIT graduate with a background in mathematics, computer science, and data science (Vance, 2021). He had no training in pandemic-related areas, such as medicine or epidemiology. He was no medical or epidemiological expert in this sense. Despite this, he saw flawed reasoning in some prominent epidemiological models, and so he wanted to try predict COVID deaths himself. He did all of this while living at his parents' house on his savings.

In time, his projections of COVID deaths outperformed many other models, such as those from the Institute for Health Metrics and Evaluation (IHME), an institute supported by over $500 million in funding. For instance, Gu successfully predicted a second large wave of infections and deaths once many US states reopened from lockdowns, while the IHME model expected the virus to wane due to social distancing and other policies. His model also predicted a month or so ahead of time that the United States would record 231,000 deaths on November 1, 2020. This prediction was highly accurate, as the United States reported 230,995 deaths on November 1, 2020. The difference between his prediction and the actual number was only five deaths.

Eventually, because of Gu's demonstrated accuracy, the Center for Disease Control and Prevention (CDC) started citing Gu's numbers on its forecasting website and participating in meetings with him.

The reason for Gu's success was that he trained his predictive models on data, developing them in a way so that they had success in accommodating old data and predicting new data about COVID deaths.

In a sense, then, he was successful because he relied on models that had a track record of accuracy, not just for accommodating the past, but also for predicting the future.

A similar moral holds for other areas in human life: even in cases where we may not have track records of an *individual's* or *group's* judgments, we may have track records for the accuracy of *models* or *other theories* which we can rely on.

This, in fact, is arguably one reason why we can be confident in much of climate science: not only have various climate models successfully accommodated past data, but for 50 years, they have successfully predicted rising temperatures as a result of growing greenhouse gas emissions (Cornwall, 2019).

Of course, not all models are so accurate, and track records are not always the only thing that matters, but the point is that much trustworthy science is built on theories or models with great track records—and rightly so.

8.2.4 Pay Attention to Qualifiers

To estimate accuracy, we often need to do something we that we do not always do: pay attention to qualifiers. By a qualifier, I mean terms that indicate a specific level of confidence: consider, for example, the expressions that "It *might* rain," "The Democrats *probably* will win," or "There is a *decent chance* that this medication is safe"; the qualifiers here are words like "might," "probably" or "decent chance," each of which indicate particular levels of confidence.

As we saw in Chap. 2, we often cannot make judgments that are fully certain, so assessing the accuracy of our less-than-certain judgments involves assessing the calibration of our levels of confidence. But we cannot assess such accuracy if we do not know what these levels of confidence are. For that reason, in assessing the accuracy of others, we must pay attention to qualifiers like "might," "probably," or "decent chance" that can signify what their levels of confidence are.

However, people often neglect to pay attention to qualifiers in daily life. Some years ago, I told some friends that I "might" do a PhD at the Australian National University. Then, weeks later, I was congratulated by other friends for my decision to do a PhD there, even though I never announced such a decision. Stories like these and countless others in my life suggest humans frequently forget or pay little attention to qualifiers, although I have not yet seen studies that show this to be the case.

Of course, one problem is that numerous studies show verbal expressions of uncertainty are ambiguous in the sense that people vary widely in their understanding of what the expressions mean (Beyth-Marom, 1982; Brun & Teigen, 1988; Clarke et al., 1992; Dhami, 2018; Dhami & Wallsten, 2005). For example, Dhami (2018) found that intelligence analysts used the word "unlikely" to represent values as low as 10% (which warrants some confidence that an event will not occur) and as high as 40% (which is closer to the probability that a coin lands heads). It is for this reason that many scholars have argued numeric probabilities should be used instead of verbal expressions of uncertainty (Dhami & Mandel, 2021; Friedman, 2019; Mandel, 2020; Mandel & Irwin, 2021a, b).

This makes it challenging to evaluate the accuracy of others when it is not precise as to what their levels of confidence are.

That said, the point of this recommendation is merely that if we want to evaluate the accuracy of people's levels of confidence, then we have to pay attention to qualifiers which indicate what those levels of confidence are in the first place—even if those qualifiers are not so precise.

8.2.5 Do Not Estimate Accuracy Based on One-Off Successes or Failures

Chapter 2 made it clear that often one-off successes or failures are not strong grounds for estimating the accuracy of others: even a perfectly accurate person will be 90% confident in things that are false 10% of the time. So, we should not rely too heavily on such successes or failures; instead, it is often only with more observations that we can estimate the accuracy of others with any confidence.

8.2.6 Do Not Always Estimate Accuracy from Years of Experience, Education, Fame, or Confidence Levels

We also should be careful about evaluating the accuracy of others using particular traditional metrics, such as how many years of experience they have in a particular area, how educated they are, how famous they are, or how confident they are in their own judgments.

As Tetlock (2005) notes in his study of political "experts," factors like these often had no correlation with accuracy:

> It made virtually no difference whether participants had doctorates, whether they were economists, political scientists, journalists, or historians, whether they had policy experience or access to classified information, or whether they had logged many or few years of experience in their chosen line of work. (p. 68)

He also noted that fame correlated negatively with accuracy, since fame can arise from inaccuracy: that is, inaccurate individuals tend to jump to conclusions with high confidence and are consequently sought out by the media for their "interesting" opinions.

Aside from that, Chaps. 3 and 4 showed study after study where people are overconfident in their own accuracy.

For this reason, these metrics of accuracy (years of experience, education, fame, or confidence levels) should be treated with caution.

That said, however, a caveat about education is needed. Intelligence correlates somewhat with accuracy (Mellers et al., 2015a), moderately with grades (Roth et al., 2015), and possibly somewhat with place of education (in virtue of more selective schools requiring higher grades). So educational achievement (qua grades) and accuracy may not be entirely irrelevant to each other, even though one may not suffice to confidently predict the other. Additionally, some kinds of education may correlate with accuracy, especially if they track the accuracy of those completing the educational programs and if their content demonstrably conduces to accurate judgments.

The problem is, however, many education programs are not like this: they neither track accuracy nor tailor their content based on what provably improves accuracy in the relevant domains. For that reason, while education may conduce to accuracy in some circumstances, there are many others where it may not.

And another caveat is needed too: while many kinds of education may not improve accuracy, that does not mean they have no value. A PhD, for example, may provide helpful skills in writing, in analyzing literature, and in understanding more about very specific domains. The point is merely that education like this may be insufficient for high levels of accuracy about open questions in these or other domains.

8.2.7 Trust Experts, But Not Too Much

In Chap. 3, we saw numerous studies indicating that experts are frequently less reliable than they or others would hope. Tetlock and Gardner (2015) also found that non-experts often did better at predicting geopolitical events than many experts. What this shows is that experts are fallible, and sometimes non-experts are even more reliable than experts.

What it does not show, however, is that non-experts are more reliable than experts on average, nor that experts are completely unreliable. Even the studies revealing the fallibility of experts indicate they still get it right to some extent. Tetlock's experts were still certain of things that were right most of the time, even if they were otherwise false in a sizable minority of cases. And no studies, so far as I can tell, have shown that non-experts are more reliable on average than experts. And in Tetlock's case, even the non-experts that were better at forecasting than experts were a rare minority of people with exceptional cognitive abilities.

What this means, I take it, is that we should trust experts, but not too much. We should have some confidence that what they say is probably true, but we should expect that they will be wrong a proportion of the time, even in cases where they are certain and do not realize their error. The fact that they are sometimes wrong means we should listen to their opinions critically, sometimes seek other opinions, and ideally look at any relevant track records of accuracy.

But of course, not all experts are equal, some fields of expertise may be more reliable than others, and if an expert has a perfect track record of accuracy, then we can trust them in the domain where they have been accurate, and perhaps others too. Yet otherwise, some degree of caution is always useful.

And since experts are fallible, this means that it is sometimes justifiable for non-experts to disagree with the opinions of some experts. Tetlock's superforecasters were sometimes non-experts who did just that, but they were provably more accurate than the experts they disagreed with.

One might object that this opens the door to undue distrust of experts—to vaccine skepticism, to climate change denialism, and to other putatively unreasonable forms of doubt.

But I do not think this is the case.

Groups of hyper-skeptics (like particular vaccine skeptics) have made mistakes, but their mistake is not simply that they do not trust experts. Experts will sometimes be wrong, in which case it is sometimes justifiable to distrust them. Instead, at most, hyper-skeptics are mistaken because they distrust genuinely good evidence, and in these cases, if the experts have accurately appraised that good evidence, then doubting them means also doubting good evidence. But the error is not distrust of experts; it is distrust of good evidence.

I think that part of the problem is that hyper-skeptics lack training in good thinking: they have not, for example, been taught to avoid confirmation bias or to appreciate the logic of scientific practices, thereby causing them to dismiss even good evidence when they see it.

Ideally, though, I think the solution to society's distrust of experts should be twofold. The first part is educating the public in good thinking, ideally in schools, so that they can discriminate between good evidence and bad evidence, between reliable experts and unreliable experts. The second part is to reform the education and ongoing recertification of experts by (i) requiring that they demonstrate track records of accuracy in their domains and by (ii) ensuring they are educated using material that demonstrably enhances accuracy in the relevant domains. Failing this, some amount of distrust of experts will not only be intellectually justifiable; it will be necessary, simply because studies of experts will continue to reveal their inaccuracy.

Aside from that, if we pretend experts are more accurate than they are, then we risk letting others rely on unreliable experts when it harms them—and maybe even kills them.

Ultimately, I think we should spend more time acknowledging inaccuracy and trying to remedy it. That way, we can reduce undue distrust of expects, not because we will blind people to the truth, but because the scientific evidence will show that experts are actually trustworthy—that they deserve the trust we want others to have. This, I take it, is better than hiding our problems under the carpet for undeserving people to blindly trip on and ultimately hurt themselves.

8.2.8 Listen to Non-Experts, But Not Uncritically

Another recommendation is to listen to the opinions of non-experts, but not uncritically.

The simple fact of the matter is that sometimes non-experts have accurate judgments, judgments with a lot of value. This was seen in the case of Youyang Gu described earlier in this chapter: despite no training in health-related fields, his models predicted COVID deaths to a degree that sometimes outperformed *all* of the alternative models developed by experts (Vance, 2021). It was also seen in the case of superforecasters: many superforecasters had no formal education in fields related to politics, but their forecasting abilities are among the best that are known in the world and far better than many political experts (Tetlock & Gardner, 2015).

It was arguably also the case in March 2020 at the beginning of the COVID pandemic. At that time, during a television debate, Silicon Valley CEO Tomas Pueyo debated an epidemiologist, with Pueyo arguing that the United Kingdom should go immediately into lockdown or else people would needlessly die. Pueyo was a Stanford-trained businessman, yet he had no training or recognized expertise in medical fields. But to justify his opinion, he appealed to statistics about prevalence rates, infection transmission rates, case fatality rates, and the experience of Italy, Spain, and China. In response, the epidemiologist dismissed his opinions, saying the United Kingdom should not go into lockdown immediately, if at all, and that the "only way to stop this epidemic is indeed to achieve herd immunity" ("Coronavirus Special: Are We Doing Enough? - Channel 4 News," 2020). The UK government listened to the epidemiologist and decided not to implement an immediate lockdown.

However, nearly 3 months later on June 6th, 2020, around 39,048 people had died from the virus, with it spreading far more quickly than the epidemiologist and others had anticipated. On another news channel, the same epidemiologist was asked if he had any regrets about the advice he gave to the UK government. He replied, "Yes… I wish we had gone into lockdown earlier. I think that has cost a lot of lives unfortunately" (*Coronavirus: Lockdown Delay "Cost a Lot of Lives", Says Science Adviser – BBC News*, 2020).

Ultimately, then, Pueyo, the non-expert, was right, whereas the expert was wrong. Now, of course, there is some debate about whether lockdowns were a good idea, but even if one thinks they were a bad idea, what is clear is that Pueyo was still more accurate than the expert was *when it came to estimating how fast the virus would spread and how deadly it was*. On that count, he was undeniably right while the expert was wrong.

Despite this, many dismissed Pueyo's opinion solely because he was a non-expert. At the time of Pueyo's debate, one person Tweeted, "I would rather put my faith in a scientific expert than a Silicon Valley businessman....really didn't understand why this overly excitable chap was on the programme" (*"We Need to Catch This before the Weekend.,"* 2020). Another Tweeted, "Please interview people who know what they're talking about, not an entrepreneur from silicon valley, with no discernible qualifications about viruses. Well done to the Prof" (*"We Need to Catch This before the Weekend.,"* 2020).

But if people had listened to Pueyo despite his lack of expertise, thousands of people may not have died and would still be alive today alongside their families who loved them.

The reason he was worth listening to, however, was because his reasoning was sound. His argument was relatively simple: the infection and fatality rates in China, Spain, and Italy warranted a lockdown, and the rates in the United Kingdom will too, at least if we want to mitigate the spread and fatality of the virus. Now, again, even if one thinks lockdowns were not the right thing to do, he still more accurately predicted the infection and fatality rates of the virus than the expert did.

For these reasons, it pays to listen critically to non-experts too, especially when they justify their opinions with good reasoning (and maybe also when they demonstrate exceptional academic achievement, since Gu and Pueyo were both educated at highly selective universities). After all, sometimes "non-experts" will get it right while the "experts" get it wrong, and when they do, ignoring non-expert opinions can be costly, sometimes deadly.

8.2.9 Beware of Negative Social Influences

Lastly, our judgments can be influenced by social forces, as discussed in Chap. 5. One of these forces is conformity, the tendency to believe what other people in our social group believe. Another force is industrial selection or other kinds of corporate influence. As we saw in that chapter, these forces can be pernicious, making us believe things that are not only false, but sometimes also deadly.

For that reason, we should not automatically assume the accuracy of judgments that profit corporations or that come from those in our social groups. We know social conformity and industrial selection can produce many false judgments, for example.

Instead, we should be aware of these influences, asking whether our judgments are produced because of good reasons or instead because we are conforming to the judgments prevalent in our society or to the judgments that are profitable for corporations to induce in us. Again, however, this does not mean a judgment is false merely because others in our group share it or because it is profitable to some corporation; it merely means we must be vigilant by paying attention to the reasons by which we arrive at our judgments.

8.2.10 Tolerate Length and Nuance

In the previous chapter, we saw more accurate forecasters had lengthier word counts for their rationales. But this was for good reasons: more accurate forecasters had lengthier rationales because they considered more competing viewpoints and issues, thus leading them to more informed and accurate rationales.

For these reasons, we should sometimes tolerate length and nuance, especially in an age where social media and other phenomena might habituate us to short clickbait and "too long; didn't read" thinking. Otherwise, we could miss out on accurate judgments merely because we impatiently dislike something that demonstrably correlates with accuracy.

8.3 Category 3: Managing Businesses or Other Organizations

So, we have examined two kinds of recommendations, those for improving our own accuracy and those for estimating the accuracy of other sources.

Lastly, I conclude with recommendations for businesses or other organizations.

8.3.1 Adopt all the Recommendations in the Previous Category

For a start, adopt the recommendations in the previous category. Organizations frequently have to estimate the accuracy of information from others. For this reason, it pays to adopt the earlier recommendations: be skeptical of others, but not too skeptical; estimate accuracy using track records where possible; look for models or theories with track records of accuracy; trust experts, but not too much; listen to non-experts, but not uncritically; and so on and so forth.

8.3.2 Promote Motivation and Accountability in Your Organization

But it may also be useful to adopt ideas from the first category of recommendations too. If your organization requires accurate judgments, promote the motivation to be accurate (either by fostering motivation or by recruiting motivated individuals), hold people accountable, and so forth.

8.3.3 Measure Track Records

In the first category, I mentioned the measurement of track records, but it is worth noting again: robust track records are reliable metrics of accuracy, but many other metrics of accuracy are not (such as fame, years of experience or even levels of confidence). For that reason, if accuracy is required in your organization, measure it. Unfortunately, however, it seems this is seldom done; according to the founder of the *Society to Improve Diagnosis in Medicine*. Mark Graber, for example, “not a single healthcare organization is measuring the diagnostic error rate” (August 11th, 2022, personal communication).

8.3.4 Give Feedback

But after measuring it, to improve accuracy, provide feedback. As mentioned, research has shown that feedback can improve accuracy in various contexts (Callender et al., 2016; Moore et al., 2017; Saenz et al., 2019).

8.3.5 Expect Backlash from the Inaccurate

Yet when giving that feedback, expect backlash, at least from some groups. As mentioned in Chaps. 3 and 4, copious amounts of evidence suggest many people are unaware of their inaccuracy and may object to feedback partly for those reasons. Additionally, as mentioned in Chap. 4, there may be strong motivational biases which incline inaccurate people to protest the metrics which reveal their inaccuracy. Recall that Tetlock's experts repeatedly objected to his feedback, complaining that they would have looked better had their accuracy been measured in other ways. While sometimes objections like this may be legitimate, they simply were not in Tetlock's case. Instead, it seems those who were the most inaccurate had the strongest motivation to protest the ways they were measured, each of which pointed to their accuracy, even after accommodating their protests and suggested changes.

8.3.6 When Possible, Create Teams, Especially of Those with the Best Track Records

Where accuracy about question is particularly important, create a team to think about that question, especially a team of those who are likely to be the most accurate. The results of Tetlock and his colleagues found this improved performance, especially when the best were matched with the best.

8.3.7 Give Training

If possible, provide training for those in your organization. Again, research shows this can improve accuracy. But again, only rely on training programs utilizing material with proven efficacy: from my analysis of the literature, much of the training that purports to improve cognitive performance actually does not.

8.3.8 Make Accuracy Profitable

My last recommendation deserves a disclaimer. Ordinarily, I would not like to talk about money in a book like this. While important, I think society often places too much value on money compared to altruism and other important values. And ideally, I would hope that people would promote accuracy solely because they recognize its value for the well-being of ourselves and others.

But for better or for worse, many people will simply not be motivated to do something unless money is involved. Fortunately, however, I think there are potentially ethical ways in which it is financially beneficial to improve accuracy so that society at large can benefit.

So, for that reason, I have a final recommendation for profit-oriented organizations in our current economies: make accuracy profitable. By that, I mean explore ways in which your business or organization can obtain an advantage over other organizations by ensuring accuracy in what you deliver. If, for example, you are a private healthcare organization, implement calibration training and measures, and then market the fact that your medical services, unlike others, are actually accurate, thereby reducing the chance of misdiagnoses, patient harm and even death. Some consumers would pay for accuracy, and people in general deserve it. For that reason, when so much of society flies blind or is inaccurate, you can offer a superior service or product for your clients and consumers by providing accuracy.

That concludes my outline of various recommendations about accuracy. These recommendations, and the content of the rest of this book, will be summarized briefly in the next chapter.

References

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes, 160*, 19–35. https://doi.org/10.1016/j.obhdp.2020.02.001

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1*(3), 257–269. https://doi.org/10.1002/for.3980010305

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes, 41*(3), 390–404. https://doi.org/10.1016/0749-5978(88)90036-2

Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning, 11*(2), 215–235. https://doi.org/10.1007/s11409-015-9142-6

Chang, W., Atanasov, P., Patil, S., Mellers, B. A., & Tetlock, P. E. (2017). Accountability and adaptive performance under uncertainty: A long-term view. *Judgment and Decision Making*, *12*(6).

Clarke, V. A., Ruffin, C. L., Hill, D. J., & Beamen, A. L. (1992). Ratings of orally presented verbal expressions of probability by a heterogeneous sample. *Journal of Applied Social Psychology, 22*(8), 638–656. https://doi.org/10.1111/j.1559-1816.1992.tb00995.x

Cornwall, W. (2019). Even 50-year-old climate models correctly predicted global warming. *Science*. https://www.science.org/content/article/even-50-year-old-climate-models-correctly-predicted-global-warming#:~:text=Most%20of%20the%20models%20accurately,today%20in%20Geophysical%20Research%20Letters%20.

Coronavirus Special: Are We Doing Enough? – Channel 4 News. (2020). *YouTube*. https://www.youtube.com/watch?v=C98FmoZVbjs&t=1162s

Coronavirus: Lockdown Delay "Cost a Lot of Lives", Says Science Adviser—BBC News. (2020, June 7). *BBC News*. https://www.bbc.com/news/uk-politics-52955034

Dhami, M. K. (2018). Towards an evidence-based approach to communicating uncertainty in intelligence analysis. *Intelligence and National Security, 33*(2), 257–272. https://doi.org/10.1080/02684527.2017.1394252

Dhami, M. K., & Mandel, D. R. (2021). Words or numbers? Communicating probability in intelligence analysis. *American Psychologist, 76*(3), 549. https://doi.org/10.1037/amp0000637

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities. *Memory & Cognition, 33*(6), 1057–1068. https://doi.org/10.3758/BF03193213

Friedman, J. A. (2019). *War and chance: Assessing uncertainty in international politics*. Oxford University Press.

Haran, U., Ritov, I., & Mellers, B. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*.

Karvetski, C. W., Meinel, C., Maxwell, D. T., Lu, Y., Mellers, B. A., & Tetlock, P. E. (2022). What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? *International Journal of Forecasting, 38*(2), 688–704. https://doi.org/10.1016/j.ijforecast.2021.09.003

Mandel, D. (2020). *Assessment and communication of uncertainty in intelligence to support decision-making*.

Mandel, D. R., & Irwin, D. (2021a). Tracking accuracy of strategic intelligence forecasts: Findings from a long-term Canadian study. *Futures & Foresight Science, 3*(3–4), e98. https://doi.org/10.1002/ffo2.98

Mandel, D. R., & Irwin, D. (2021b). Uncertainty, intelligence, and National Security Decisionmaking. *International Journal of Intelligence and CounterIntelligence, 34*(3), 558–582. https://doi.org/10.1080/08850607.2020.1809056

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., Tetlock, P. E., Stone, E., & Tetlock, P. E. B. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015a). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1–14. https://doi.org/10.1037/xap0000040

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015b). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science, 10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Mellers, B. A., Tetlock, P. E., Baker, J. D., Friedman, J. A., & Zeckhauser, R. (2019). Improving the accuracy of geopolitical risk assessments. In H. Kunreuther, R. J. Meyer, & E. O. Michel-Kerjan (Eds.), *The future of risk management* (pp. 209–226). University of Pennsylvania Press.

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science, 63*(11), 3552. https://doi.org/10.1287/mnsc.2016.2525

Roth, B., Nicolas, B., Sara, R., Sarah, S., Florian, D., & Frank, M. S. (2015) "Intelligence and school grades: A meta-analysis." *Intelligence, 53*: 118–137.

Saenz, G. D., Geraci, L., & Tirso, R. (2019). Improving metacognition: A comparison of interventions. *Applied Cognitive Psychology, 33*(5), 918–929. https://doi.org/10.1002/acp.3556

Tetlock, P. (2005). Expert political judgment: How good is it? How can we know? Princeton University Press.

Tetlock, P., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Broadway Books. https://doi.org/10.1201/b15410-25

Vance, A. (2021). The 27-year-old who became a Covid-19 data superstar. *Bloomberg Businessweek*. https://www.bloomberg.com/news/articles/2021-02-19/covid-pandemic-how-youyang-gu-used-ai-and-data-to-make-most-accurate-prediction.

We Need to Catch This Before the Weekend. (2020, March 13). *Twitter*. https://twitter.com/channel4news/status/1238573667002523648

Wilcox, J. (Working paper A). Credences and trustworthiness: A calibrationist account.

Wilcox, J. (Working paper B). *Measures of accuracy*.

Chapter 9
Conclusion

This book was, I hope, an appropriate mixture of ups and downs, goods and bads.

The bad news is about how accurate we are. According to the *context-dependent model* of human accuracy, we are accurate in some cases, worryingly inaccurate in others and then, in many other contexts, we simply do not know—we fly blind and rely on the accuracy of ourselves and others in ways which are not currently warranted by the evidence. The result of judgmental inaccuracy is often extreme, compromising our decisions and even costing thousands of lives across areas like medicine, law, and politics. Unfortunately, our inaccuracy is often undetected, and there are various explanations for why this is the case: because the skills that are necessary to form accurate judgments are the same that are necessary to recognize our own inaccuracy, or because we do not track the accuracy of our own judgments, for example. Additionally, a range of heuristics can explain how we arrive at judgments that are often accurate and adequate, but also often not. All of this may be expectable given our evolutionary history, either because evolution permits functional defects in our truth-seeking abilities or because such "defects" expediently conduce to other more social functions.

In any case, the literature is depressing at times but hopefully in a useful way: it can alert us to areas to improve our own judgments, thereby helping us to make better decisions and live better lives.

The more positive message of this book, then, is that we can be accurate at times and—most importantly—we can improve our accuracy. Insights from empirical epistemology can tell us how to do this.

Consequently, in the preceding chapter, I outlined some recommendations about the accuracy of human judgment according to the relevant categories below:

Recommendations for Improving Our Own Judgments

1. Foster motivation
2. Become accountable
3. Track your accuracy

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0_9

4. Be your own skeptic: expect inaccuracy and embrace humility
5. Beware of intuition
6. Practice active open-minded thinking
7. Gather subject-specific knowledge and from diverse sources
8. Use statistics, especially base rates
9. Average estimates from conflicting sources
10. Test for scope sensitivity
11. Do postmortems
12. Take some training

Recommendations for Estimating the Accuracy of Other Sources

1. Be skeptical of judgment, but not too skeptical
2. Estimate accuracy based on track records
3. Look for models or theories with track records of accuracy
4. Pay attention to qualifiers
5. Do not estimate accuracy based on one-off successes or failures
6. Do not always estimate accuracy from years of experience, education, fame, or confidence levels
7. Trust experts, but not too much
8. Listen to non-experts, but not uncritically
9. Beware of negative social influences
10. Tolerate length and nuance

Recommendations for Managing Businesses or Other Organizations

1. Adopt all the recommendations in the previous category
2. Promote motivation and accountability in your organization
3. Measure track records
4. Give feedback
5. Expect backlash from the inaccurate
6. When possible, create teams, especially of those with the best track records
7. Give training
8. Make accuracy profitable

I think that, if implemented, these recommendations could substantially improve the accuracy of our judgments.

However, as research continues in this area, these sets of recommendations may change or become more refined. For that reason, interested readers may wish to follow my website and social media account(s) for updates and more about judgmental accuracy: visit https://johnwilcox.org/.

In any case, while many studies paint a less-than-ideal view of human judgment, my aspiration is that these recommendations can help us move closer to the ideal—that is, closer to accurate judgments. As a result, hopefully we can apply the insights from empirical epistemology in ways that improve our judgments, our decision-making and ultimately our lives—both as individuals and a collective.

Appendix: Judgments and Emotions

This book has focused on human judgment, how accurate it is, and how to improve it.

We saw that, in many contexts, humans are often much more inaccurate than we would hope or expect.

This, then, has implications for a topic that is important for us all: our emotions. Here, we will work within the "feeling tradition" of emotions, understanding "emotions" to refer to feelings like love, hate, disgust, sadness, anger, and coldness—as well as other states such as desires, urges, and the like (Scarantino & de Sousa, 2018). (Psychologists also often distinguish between "moods" and "emotions" too, but we can regard them as the same for our purposes here.)

In this appendix, we will explore some ideas about our emotions that are enshrined in cognitive behavioral therapy, ideas that make a close connection between judgments and emotions. Since this book is about the accuracy of human judgment, we will also consider some of this book's implications for the topic of our emotions, and we will finally explore some ideas about how to approach and respond to our emotions.

The Close Connection Between Judgment and Emotions

Many examples can illustrate the common-sense connection between our emotions or feelings on the one hand and our judgments or beliefs on the other. Imagine you receive a rejection for a job application, so you come to believe you will never get a good job, and this causes feelings of sadness. But then you later learn one of your job applications was successful, and you start to feel happy since you now believe

Disclaimer: Note the following content is not intended as medical advice to any readers. Readers are perusing the following content at their own risk and should consult a professional healthcare provider in treating any conditions they have.

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0

you have been offered a good job. Or imagine you see a beautiful bouquet of flowers arrive at your neighbor's door, and you feel resentment toward your partner since you believe they were not considerate enough to buy you flowers like your neighbor's partner. But then you learn that—to your delight—the flowers were misdelivered and were actually intended for you. So your feelings of resentment transmute into feelings of love for your partner. Or imagine you feel awful anxiety after receiving threatening emails from a stranger and then believing your life is in danger. But then you feel happiness and amusement when you come to realize the emails were a prank from a friend.

In each of these cases, the emotions are not caused directly by reality or by the events themselves; the emotions are caused by the *judgments*, *beliefs* or *cognitions* about reality. (For simplicity's sake, we can think of "judgments", "beliefs" and "cognitions" as referring to more or less the same thing in this appendix.) The sadness was caused by the judgment that you will never get a good job, even though, in reality, this judgment was false. The resentment was caused by the judgment that your partner is not as considerate as your neighbor's partner, even though, in reality, this judgment was false. The anxiety was caused by the judgment that your life is in danger, even though, in reality, this judgment was false.

We can also imagine other cases where virtually *any* emotion or feeling is produced by false judgments, including happiness, desires, urges, feelings of "needs", and so forth.

Of course, often we are not mistaken about our emotions or the judgments that produce them.

Nevertheless, the cases where we are mistaken clearly illustrate the central role that judgments—not reality—can play in directly producing our emotions.

Furthermore, what these examples illustrate is that these emotions or feelings can *change* when the underlying judgments themselves change. You became happy once you believed you were offered a good job; you felt love once you believed your partner was considerate; and you felt amusement once you believed your friends played a humorous prank on you.

These examples also illustrate that insofar as our judgments can be mistaken, so too can the emotions that arise from them also be mistaken. Above, the feelings of sadness, resentment, and anxiety were all, in a sense, *epistemically* mistaken insofar as they were based on *inaccurate judgments* about the reality. This is true even if they are valid in other respects, such being based on a reasonable desire for a job or for a considerate partner.

Additionally, not only can these emotions be mistaken, but they can be *maladaptive*: that is, they can compromise our well-being, either directly by negatively affecting how we feel at the time or indirectly by leading to actions that compromise our well-being later. For example, the resentment you felt might be the final straw in a series of emotions, which themselves may be mistaken, and this may make you rashly break up with your partner and potentially make you feel worse in the long term than had the emotions been reappraised.

Cognitive Behavioral Therapy

In a sense, all these ideas are enshrined in *cognitive behavioral therapy*, an evidence-based treatment for a range of emotional conditions. The central premise of cognitive behavioral therapy is that emotions and disorders like anxiety and depression arise not directly from reality itself, but rather from our *judgments* about reality—judgments about ourselves, others, the future, and the world more generally (Hofmann et al., 2012; Kazantzis et al., 2018). Cognitive behavioral therapy also holds that these judgments are sometimes maladaptive in ways that compromise our well-being. Consequently, cognitive behavioral therapies to treat these emotional disorders focus on understanding, challenging, and replacing the maladaptive judgments that produce these emotions. That said, although the role of judgments or cognitive factors is recognized, attention is also given to physiological and behavioral components that contribute to disorders (Hofmann et al., 2012); it is clear that sometimes hormonal or other physiological changes can sometimes themselves affect emotions or moods, regardless of the judgments one holds.

So far, a large body of evidence has reported that cognitive behavioral therapy is successful in treating a range of conditions ranging from depression and anxiety to excessive anger and bulimia (Hofmann et al., 2012; Kazantzis et al., 2018).

Ultimately, then, there is plausibly a strong connection between judgments on the one hand and emotions on the other hand.

If this is the case, then the content presented in this book has some implications for our emotions.

We have seen that our judgments about reality are often more inaccurate than is generally appreciated, and we know that judgments often produce emotions.

If this is the case, then our emotions may often be more mistaken than we realize, since they may be produced by inaccurate judgments about the world.

For that reason, if the judgments producing our emotions are more mistaken than we may realize, clearly it is inappropriate to place unquestioning trust in our own emotions or to unhesitatingly validate the legitimacy of other people's emotions.

Yet this idea conflicts with particular common societal attitudes in two ways.

First: put simply, people often do not like to have their feelings invalidated—to feel that they are wrong or to be told that they are wrong. This has two consequences. First, many people seldom question their own emotions, because acknowledging wrongness can be painful and saddening in itself. Second, many people sometimes do not question the emotions of others—at least not directly to them—especially since this can create interpersonal conflict, dislike, and distance. Often, it is easier to avoid these conflicts by validating the feelings of others. This is true of friends who may wish to validate each other's feelings, but it is also true of, say, businesses who may wish to validate the feelings of their customers.

This unquestioning validation of emotions is simply naïve. While we must accept the fact that we all experience emotions, we must also accept the fact that our emotions are sometimes—if not often—based not on facts, but rather on false judgments. To think otherwise is simply to ignore the truth, to ignore the science of human judgment, and to ignore the obvious fact that all humans are fallible.

Second: there is a widespread dichotomy between the "head" and the "heart", between thinking and feeling. More specifically, attempts to evaluate emotions with facts are just seen as taking a different and inferior mode of dealing with emotions—the mode of "thinking" rather than "feeling". As a result, discussions about emotions become unchallengeable and impervious to reality.

Yet this distinction between head and heart—between thinking and feeling—is also mistaken. Feelings affect thoughts, and thoughts affect feelings; that is, the way we think can produce and affect the way we feel, and vice versa. Many of our most successful approaches for treating emotional disorders rest precisely on this idea. It is then naïve to say that successfully managing emotions via understanding and challenging our thoughts somehow neglects feelings. To successfully understand and manage our feelings, we need to think about them and their causes.

So this book, coupled with insights from cognitive behavioral therapy, supports the fallibility of emotion.

One might think that the appropriate response to our fallibility is simply to ignore or suppress the expression of our emotions, especially if they may be so mistaken. This response is known as "emotional suppression" or "expressive suppression".

The available evidence suggests, however, that this response itself is mistaken.

Multiple studies indicate that, compared to people who do not suppress the expression of their emotions, those who do suppress their emotions are less likely to experience positive emotions and more likely to experience negative emotions, including feelings of inauthenticity and depressive symptoms (Gross, 2014; Gross & John, 2003; Moore et al., 2008; Nezlek & Kuppens, 2008). The evidence also indicates expressive suppression is more likely to negatively affect one's relationships with others in various ways, such as having less positive relations with others and less liking from partners in social interactions (Butler et al., 2003; English et al., 2013; Gross & John, 2003; Srivastava et al., 2009).

So the evidence does not favor suppressing the expression of our emotions, although it may sometimes be appropriate to *temporarily* suppress the expression of emotions until they can be appropriately expressed.

Ignoring our emotions is also sometimes inappropriate for another reason: although our emotions are sometimes based on mistaken judgments, the rest of the time, they are not. Sometimes they are trustworthy, and sometimes they are not. So the results presented in this book do not justify full-blown distrust of emotion either.

How, then, should we approach our emotions? In particular, how can we tell whether we place too much or too little trust in our emotions and in the emotions of others?

The findings presented in this book would suggest a two-step process—one that I think is likely reflected in many cognitive behavioral therapy interventions.

Step One: Understanding Emotions Via Understanding Their Underlying Judgments

The first step is to *understand* what emotions we are feeling and *why* we are feeling them—or, more specifically, to understand any *judgments* that are producing our emotions. For example, suppose you are feeling angry toward a friend or an urge to break up with your partner. The first step would be to understand *why* you feel this way: for example, perhaps you are angry because you judge that your friend lied to you about something that was important to you, or perhaps you feel an urge to leave your partner because you judge that you have been arguing too much recently and would be happier without them. The idea is then to first understand our emotions by understanding the underlying judgments.

People vary in their ability to understand the judgments that produce their emotions. Studies suggest those who are better at differentiating and understanding their emotions are more likely to experience a range of positive outcomes, such as higher self-esteem, less depression, less neuroticism, and more adaptive emotion regulation (Barrett et al., 2001; Demiralp et al., 2012; Erbas et al., 2014; Schwarz, 1990).

Understanding one's emotions is also one of four components in how emotional intelligence is traditionally understood, alongside the abilities to perceive emotions, to use emotions to facilitate "thought", and to regulate emotions (Brackett et al., 2016).

Fortunately, it seems some interventions may be able to improve both emotional intelligence in general and the ability to understand emotions in particular.

One approach is the practice of *expressing* one's emotions. This can look like talking with others, such as friends, family, or a therapist. It can also look like so-called *expressive writing*—that is, "the procedure of writing freely and emotionally about a personal topic or event without paying attention to grammar or spelling" (Reinhold et al., 2018, p. 1). This basic idea is also called *written emotional disclosure* (Frisina et al., 2004), *scriptotherapy* (Riordan, 1996), *therapeutic writing* (Wright & Chung, 2001), or *self-reflective journaling* (Kremenitzer, 2005). Various studies and meta-analyses have found that expressing oneself—either verbally with others or through expressive writing—is associated with a range of positive psychological and physiological health outcomes (Frattaroli, 2006; Frisina et al., 2004; Smyth, 1998), such as reduced hostility (Austenfeld et al., 2006) and depressive symptoms (Pope et al., 2006).

So a range of studies indicate that expressing oneself can help us to understand and manage our emotions, but there is also anecdotal support for this too. For example, Janet Kremenitzer trains early childhood educators—an often emotionally demanding profession. She recommends reflective journaling as an important way to develop emotional intelligence. She says her students consider journaling to be an "essential part of their 'bag of tricks' that they need to have as they begin their careers" (Kremenitzer, 2005, p. 5). She also provides a "typical reflection" from an educator about the sometimes surprisingly helpful effect of journaling and writing about one's emotions:

> I have to admit that at first I did not think that keeping a journal to track your emotional intelligence was going to be an effective method for me. I was very wrong. I have learned so much about myself throughout this process. I feel that I have in fact developed a keener sense of "hyper-awareness". This is an essential skill for all teachers to possess. I find that teaching (sic) not only can it be very stressful and busy, but at times, it can be emotionally taxing. By becoming aware of who you are and how your (sic) react, you are better able to become a competent and professional educator. I find that even in my personal life, remembering to "catch myself" has saved me a lot of grief in the end. After spending a significant amount of time trying to become aware of my emotional self, I realize that most issues both professionally and personally are better resolved when you are truly in a calmer place, such as neutral. To give your self a moment to consciously make the choice, to get to a different place emotionally, is the key to productive problem solving. My secret weapon is now being able to control my emotional response by flipping the switch to neutral and handling a situation in a coherent and professional manner. (Student teacher. Quoted in Kremenitzer (2005, p. 5))

So there is some evidence for the benefits of expressive writing.

Yet these benefits may also depend on *who* is writing. More specifically, one meta-analysis of emotional writing among cancer patients indicated that patients with lower levels of emotional support from others may be more likely to benefit from expressive writing, while other cancer patients with greater emotional support may not benefit as much or even at all (Zachariae & O'Toole, 2015). Additionally, studies with a greater proportion of males have reported greater effects of emotional writing; some think this is because, compared to females, males are less likely to express their emotions with others (Frattaroli, 2006; Smyth, 1998). The combination of these insights suggests that those who will benefit from expressive writing are those who do not express their emotions in other ways.

The benefits of expressive writing may also depend on *how* the writing is done. One meta-analysis failed to find evidence for the efficacy of brief, self-directed writing interventions for treating depressive symptoms; however, greater effect sizes were demonstrated when, in the authors' words, "the number of sessions was higher and when the writing topic was more specific" (Reinhold et al., 2018, p. 1). More specifically, they suggest future research should involve longer sessions where "participants should write for more than three sessions about one specific topic" (Reinhold et al., 2018, p. 10). Perhaps this suggests that the topic question should often be, "What am I feeling, and why do I feel this way?".

In any case, there is some evidence for the benefits of emotional expression in enhancing self-understanding, although the benefits of expressive writing may depend on who does it and how it is done.

Ultimately, though, since some emotions rest on mistaken judgments while others do not, it seems appropriate to first try to understand the judgments that produce our emotions, using techniques such as expressive writing or communication with others.

Step Two: Challenging Judgments

Having identified the judgments producing our emotions, the second step is to challenge them and, where appropriate, replace them with positive but realistic counter-judgments, a strategy described by Mininni (2005). This can involve asking various questions. Are these judgments true? Can I get more information about whether these judgments are true? Do these judgments ignore other facts about the situation which undermine their validity? Do these facts suggest that one should adopt "counter-judgments" which are more positive and realistic?

For example, if someone is sad because they judge they will not get a good job, they can challenge this judgment, asking whether it really is the case that they will not get a good job or whether they might indeed have a chance, perhaps after more applications or after further up-skilling. Likewise, if someone is angry because they judge that their friend made an important mistake, they can challenge the judgment that the mistake is that important.

So the second step is to challenge the judgments that produce our emotions.

If the judgments are inaccurate—as they may often be—then they can be replaced with more realistic judgments, and perhaps the negative emotions will abate. When we make ourselves see the situation differently to manage our emotions, we adopt a strategy of emotion regulation called *emotional reappraisal*. This strategy is associated with a variety of benefits, such as less negative emotional experiences and closer relationships with others (Gross, 2014).

However, if the judgments producing our emotions are true, then they can provide insight into how we should act. For example, if someone is anxious because they judge some threatening event will occur, and if this judgment is accurate, then their anxiety can usefully motivate them to act in ways to reduce or cope with the threat. Psychologist Nico Frijda (1994) goes so far as to claim that the very function of emotion is to motivate behavior to cope with the emotional events.

In a sense, then, we might think of our emotions as being like messages which arrive in our mental inbox. If we choose to ignore and suppress them, studies suggest this can lead to a variety of negative outcomes. But if we choose to open and explore them, then they can teach us about ourselves and about the world; they can serve the function of alerting us to what we care about and to how we see the world—to how we judge it to be. In cases where the judgments are inaccurate, they can be replaced with more realistic judgments. But in cases where the judgments are accurate, they can often motivate useful behaviors to cope with the emotional events.

But in any case, the evidence for the fallibility of human judgment also provides evidence for the fallibility of human emotions. The aforementioned steps then constitute one way of approaching our emotions in light of this evidence—hopefully one that, again, can improve our judgments, our decision-making, and ultimately our lives.

References

Austenfeld, J. L., Paolo, A. M., & Stanton, A. L. (2006). Effects of writing about emotions versus goals on psychological and physical health among third-year medical students. *Journal of Personality, 74*(1), 267–286.

Barrett, L. F., Gross, J., Christensen, T. C., & Benvenuto, M. (2001). Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition & Emotion, 15*(6), 713–724.

Brackett, M. A., Rivers, S. R., Bertolli, M. C., & Salovey, P. (2016). Emotional intelligence. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (4th ed., pp. 513–531). The Guilford Press.

Butler, E. A., Egloff, B., Wlhelm, F. H., Smith, N. C., Erickson, E. A., & Gross, J. J. (2003). The social consequences of expressive suppression. *Emotion, 3*(1), 48.

Demiralp, E., Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Barrett, L. F., Ellsworth, P. C., Demiralp, M., Hernandez-Garcia, L., & Deldin, P. J. (2012). Feeling blue or turquoise? Emotional differentiation in major depressive disorder. *Psychological Science, 23*(11), 1410–1416.

English, T., John, O. P., & Gross, J. J. (2013). *Emotion regulation in close relationships.*

Erbas, Y., Ceulemans, E., Lee Pe, M., Koval, P., & Kuppens, P. (2014). Negative emotion differentiation: Its personality and Well-being correlates and a comparison of different assessment methods. *Cognition and Emotion, 28*(7), 1196–1213.

Frattaroli, J. (2006). Experimental disclosure and its moderators: A meta-analysis. *Psychological Bulletin, 132*(6), 823.

Frijda, N. H. (1994). Emotions are functional, most of the time. In P. Ekman & R. J. Davidson (Eds.), *The nature of emotion: Fundamental questions* (pp. 112–122). Oxford University Press.

Frisina, P. G., Borod, J. C., & Lepore, S. J. (2004). A meta-analysis of the effects of written emotional disclosure on the health outcomes of clinical populations. *The Journal of Nervous and Mental Disease, 192*(9), 629–634.

Gross, J. J. (2014). Emotion regulation: Conceptual and empirical foundations. In *Handbook of emotion regulation* (2nd ed., pp. 3–20). The Guilford Press.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and Well-being. *Journal of Personality and Social Psychology, 85*(2), 348–362. https://doi.org/10.1037/0022-3514.85.2.348

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research, 36*(5), 427–440. https://doi.org/10.1007/s10608-012-9476-1

Kazantzis, N., Luong, H. K., Usatoff, A. S., Impala, T., Yew, R. Y., & Hofmann, S. G. (2018). The processes of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research, 42*(4), 349–357. https://doi.org/10.1007/s10608-018-9920-y

Kremenitzer, J. P. (2005). The emotionally intelligent early childhood educator: Self-reflective journaling. *Early Childhood Education Journal, 33*(1), 3–9. https://doi.org/10.1007/s10643-005-0014-6

Mininni, D. (2005). The emotional toolkit: Seven power-skills to nail your bad feelings. Macmillan.

Moore, S. A., Zoellner, L. A., & Mollenholt, N. (2008). Are expressive suppression and cognitive reappraisal associated with stress-related symptoms? *Behaviour Research and Therapy, 46*(9), 993–1000. https://doi.org/10.1016/j.brat.2008.05.001

Nezlek, J. B., & Kuppens, P. (2008). Regulating positive and negative emotions in daily life. *Journal of Personality, 76*(3), 561–580. https://doi.org/10.1111/j.1467-6494.2008.00496.x

Pope, H., Watkins, K. W., Evans, A. E., & Hess, P. (2006). The perception of depression in long-term-care residents: A qualitative study using residential journaling. *Journal of Applied Gerontology, 25*(2), 153–172.

Reinhold, M., Bürkner, P.-C., & Holling, H. (2018). Effects of expressive writing on depressive symptoms—A meta-analysis. *Clinical Psychology: Science and Practice, 25*(1), e12224. https://doi.org/10.1111/cpsp.12224

Riordan, R. J. (1996). Scriptotherapy: Therapeutic writing as a counseling adjunct. *Journal of Counseling & Development, 74*(3), 263–269.

Schwarz, N. (1990). *Feelings as information: Informational and motivational functions of affective states*. The Guilford Press.

Smyth, J. M. (1998). Written emotional expression: Effect sizes, outcome types, and moderating variables. *Journal of Consulting and Clinical Psychology, 66*(1), 174.

Srivastava, S., Tamir, M., McGonigal, K. M., John, O. P., & Gross, J. J. (2009). The social costs of emotional suppression: A prospective study of the transition to college. *Journal of Personality and Social Psychology, 96*(4), 883.

Wright, J., & Chung, M. C. (2001). Mastery or mystery? Therapeutic writing: A review of the literature. *British Journal of Guidance & Counselling, 29*(3), 277–291.

Zachariae, R., & O'Toole, M. S. (2015). The effect of expressive writing intervention on psychological and physical health outcomes in cancer patients—A systematic review and meta-analysis. *Psycho-Oncology, 24*(11), 1349–1359. https://doi.org/10.1002/pon.3802

Index

J. Wilcox, *Human Judgment*, SpringerBriefs in Psychology,
https://doi.org/10.1007/978-3-031-19205-0