Reason, Bias, and Inquiry

Reason, Bias, and Inquiry

The Crossroads of Epistemology and Psychology

Edited by

NATHAN BALLANTYNE AND DAVID DUNNING

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

CIP data is on file at the Library of Congress

ISBN 978–0–19–763691–6

DOI: 10.1093/oso/9780197636916.001.0001

1 3 5 7 9 8 6 4 2

Printed by Integrated Books International, United States of America

Acknowledgments

The co-editors would like to thank the authors for their thoughtful contributions.

N. B. and D. D. would like to mention the helpful assistance of several graduate students in completing this volume: Noah Hahn, Johnny Brennan, and Christian Emrich.

N. B. received generous support from the John Templeton Foundation (grant #54160) to host an epistemology and psychology conference, held in New York City in June 2016. He would like to thank Alexander Arnold, John Churchill, and Michael Murray. N. B. expresses his gratitude to Shane Wilkins for co-organizing and co-hosting the conference.

The co-editors would like to thank Matthew Ballantyne for proposing the collection's cover image: a detail from "Planned obsolescence," an artwork featured in Nicolás Lamas's show, *Archaeology of Darkness*, at Meessen De Clercq in Brussels, Belgium in 2019. The co-editors are grateful to Nicolás Lamas for permitting use of the image and to Nick Lowen for developing the cover's design.

The co-editors are grateful to their editorial team at Oxford University Press. The collection was initially picked up by Joan Bossert and Phil Velinov, who then moved into new roles elsewhere. Abby Gross and Katie Pratt stepped in and pushed things forward, before Abby took up a new role at the press. Finally, Nadina Persaud and Katie, with production assistance from Prabha Karunakaran at Newgen KnowledgeWorks, brought everything to a great conclusion.

Contributors

Teresa Allen, PhD
Adjunct Associate Professor
Department of Philosophy
North Central College
Naperville, IL, USA

Emily Balcetis, PhD
Associate Professor
Department of Psychology
New York University
New York, NY, USA

Nathan Ballantyne, PhD
Associate Professor
Department of Philosophy
Fordham University
New York, NY, USA

Nathan N. Cheek, PhD
Postdoctoral Research Associate
Department of Psychology
Princeton University
Princeton, NJ, USA

Jason D'Cruz, PhD
Associate Professor
Department of Philosophy
University at Albany, SUNY
Albany, NY, USA

David Dunning, PhD
Professor
Department of Psychology
University of Michigan
Ann Arbor, MI, USA

Yael Granot, PhD
Assistant Professor
Department of Psychology
Smith College
Northampton, MA, USA

Kalypso Iordanou, PhD
Associate Professor
School of Sciences
University of Central Lancashire,
Cyprus Campus (UCLan Cyprus)
Larnaca, CY

Kristyn A. Jones, PhD
Research Fellow
Department of Psychology
Stanford University
Palo Alto, CA, USA

Frank C. Keil, PhD
Professor
Department of Psychology
Yale University
New Haven, CT, USA

Thomas Kelly, PhD
Professor
Department of Philosophy
Princeton University
Princeton, NJ, USA

Hilary Kornblith, PhD
Distinguished Professor
Department of Philosophy
University of Massachusetts, Amherst
Amherst, MA, USA

Deanna Kuhn, PhD
Professor
Teachers College Columbia University
New York, NY, USA

Kristi L. Lockhart, PhD
Senior Lecturer, Emeritus
Department of Psychology
Yale University
New Haven, CT, USA

Michael Patrick Lynch, PhD
Board of Trustees Distinguished Professor
Department of Philosophy
University of Connecticut Humanities Institute
Storrs, CT, USA

Sarah McGrath, PhD
Professor
Department of Philosophy
Princeton University
Princeton, NJ, USA

Hugo Mercier, PhD
Institut Jean Nicod
Département d'études cognitives
ENS, EHESS, PSL University, CNRS
Paris, FR

Jessie Munton, PhD
Associate Professor
Faculty of Philosophy
University of Cambridge
Cambridge, UK

Emily Pronin, PhD
Associate Professor
Department of Psychology, and School of Public & International Affairs
Princeton University
Princeton, NJ, USA

Roy Sorensen, PhD
Professor
Department of Philosophy
University of Texas at Austin
Austin, TX, USA

Alessandra Tanesini, PhD
Professor
School of English, Communication and Philosophy
Cardiff University
Cardiff, UK

Chris Tucker, PhD
Associate Professor
Department of Philosophy
William & Mary
Williamsburg, VA, USA

Introduction

> Epistemology without psychology is like a dinner without food, pure tastes arranged in a series of sweets, salts, bitters, and acids.
>
> —George Boas, "Philosophy and Ritual," *Proceedings and Addresses of the American Philosophical Association* (1951–1952)

Questions surrounding reasoning, inquiry, and bias are among the most enduring in human history. Ideas and theories about human reasoning and knowledge can be found in ancient philosophical writings, from Greece to China. Yet, to a great extent, these questions have never been more pressing—and unsettled—as they are today in our information-drenched contemporary society.

Thus, it is no surprise that questions about reasoning, inquiry, and bias are examined in active and emerging discussions in the contemporary fields of epistemology and psychology. But, despite their shared thematic interests, researchers in epistemology and psychology do not often talk together, read each other's writings, or collaborate. This volume arises from the conviction that the separation of the two fields is both unfortunate and unnecessary. There are opportunities for real learning between the two fields.

Consider a little thought experiment. Imagine a researcher describes their work in terms of three keywords: *knowledge*, *judgment*, *overconfidence*. Or suppose another researcher uses a different triad of terms: *irrationality*, *biases*, *expertise*. On the basis of such keywords alone, can you know whether a researcher hails from psychology or philosophy? We think not. There are practitioners in both fields who would naturally characterize their work along those lines. However, the two groups of scholars remain separated by their respective disciplinary boundaries.

To be sure, the two fields do have real differences. Epistemologists dwell on normative matters, whereas psychologists focus on descriptive ones. Epistemologists often theorize in view of what is conceivable and broadly

Introduction In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press.
 DOI: 10.1093/oso/9780197636916.003.0001

possible, whereas psychologists tend to be constrained by what is real and measurable. Despite this, they often share intellectual curiosity about the same phenomenon. And they can bring complementary insights to inquiries—ones that inform across disciplinary lines and can combine to educate the non-specialist public.

We offer this volume as a hint of fascinating and productive conversations that can happen when epistemologists and psychologists come together. Its chapters examine contemporary perspectives on reason, its pursuit (i.e., inquiry), and its potential obstacles (i.e., bias). No volume of its length can offer a comprehensive survey, but we bring to you ideas and arguments from one group of active researchers on these topics. The volume is structured around engagements between philosophers and psychologists, who attempt to characterize reason, inquiry, and bias as well as to describe ordinary people engaging in reasoning, inquiry, and bias.

As we began by noting, these issues are particularly significant in the present moment. Although one can debate whether reason is the primary element needed to survive modern challenges, we believe our era reveals that our collective surviving and thriving will at least be impossible without it. A tremendous flood of information and argument is available at the click of a mouse or swipe of a screen, and the ability to reason well about what we read, see, and hear is indispensable.

Many challenges underline just how important good reasoning has become; we will limit ourselves to three examples. Consider first the tremendous information available to governments and citizens through networked computers. Much of it is accurate and valuable, but the internet is rife with misleading information and outright lies. The World Economic Forum in 2013 cited internet misinformation and its potential to produce "digital wildfires" as a threat to human survival, akin to the threat of antibiotic-resistant germs. So much dangerous chaff is mixed in with the wheat, and it takes deliberation and discernment to separate the two.

A second challenge is the rise of political polarization, which includes the politicization of science. As Theresa Allen and Michael Patrick Lynch in this volume point out, successful democracy requires citizens who can reason together with empathy and humility. Some shared understanding of the basic facts on the ground is also important for functional civic debate. But recent surveys indicate that not only are Americans sorting themselves into more extreme camps, measured by discordant attitudes, but they are also doing so emotionally. They do not merely disagree on the issues—they also express

sheer displeasure with each other (Iyengar et al., 2019). For example, one series of Pew surveys has tracked how much Americans view their political opponents with anger and fear, describing their opponents as threats to the country (Pew Research Center, 2016, 2019). The numbers have been on the rise for several years. Although this sharpening of affective polarization might be particularly pronounced in the United States, and less so elsewhere, some other nations are on a similar upward path, including Canada, New Zealand, and Switzerland (Boxell et al., 2020).

A third challenge arose late in 2019 when a novel virus began infecting communities around the world. The virus was quickly identified as a coronavirus (and given the formal name of SARS-CoV-2, although COVID-19 became the more popular designation), but beyond that, governments, health authorities, and citizens were at a near dead stop of knowledge about how the virus wreaks havoc and how it can be combatted. What effects does it have on the body? How does one inhibit transmission? How can it be treated? Can a vaccine be developed against it, and how can vaccines be effectively distributed? The world over, researchers rushed into a battle of crisis inquiry to combat not just the virus but the sheer uncertainty surrounding how to fight it. Here, the importance of accurate reasoning and efficient inquiry came into sharp relief. The importance of getting it right was underlined by a steadily rising death toll and dire economic consequences. At the time of writing, the ultimate outcome is unsettled, but broad questions loom about the effectiveness and crisis-readiness of certain fields of knowledge as well as the public's understanding of expertise and science. A global health crisis helped to reveal an epistemological one.

The Volume's Organization

The volume is divided into three sections. In the first, "Rationality and Bias," our authors explore the conceptual and empirical geography of reason and bias. These two concepts are rich and complex, having been understood in a variety of ways by theorists and experimentalists. Philosophers Thomas Kelly and Sarah McGrath begin their chapter by noting ways to attribute bias to people, groups, opinions, objects, and outcomes. Commonplace talk of "bias" has an untidy sort of disunity, and Kelly and McGrath raise an important question: Is there one type of fact about bias that is fundamental, in the sense that other facts can be explained in terms of it? Facts about bias, they

argue, are typically based in or grounded upon facts about biased processes and procedures.

In the next chapter, Nathan N. Cheek and Emily Pronin describe the psychology behind asymmetric assignments of bias to self and others. They survey the empirical literature to show that people have a "bias blind spot," wherein people fail to recognize their own biases and so view others as the biased ones. Roy Sorensen approaches similar issues arising from how we evaluate others' rationality. We presume other people have a logic behind their judgments and actions but also admit that logic is in short supply in human cognition, in virtue of limits on attention and processing power. Thus, we both lean on rationality to understand others and then minimize rationality because it is cognitively expensive. What explains this give and take? Sorensen examines a series of philosophical perplexities arising from apparently conflicting attributions of rationality. Teresa Allen and Michael Lynch next consider the kind of reasoning required in a civil and functioning democracy, and Jason D'Cruz asks whether the reasoning contained in after-the-fact rationalizations is authentic or just mere pretense. According to D'Cruz's account, rationalizers are at bottom storytellers aiming mainly to self-justify.

In the volume's second section, "Perception and Attention," our authors examine how reasoning and bias can extend down into perceptual experience—what we literally see, hear, taste, and smell. Normally, people assume they see the physical world as it really is, but psychological research calls that assumption into doubt. Perceiving the outside world is an inherently difficult task. Our eyes and ears receive information that is often ambiguous or too sparse to guide judgment—and so the brain needs to reach conclusions in the absence of adequate data. For instance, upon viewing a hill, our eyes receive insufficient information to assess the steepness of its slope. Yet people feel they can accurately gauge that slope. Furthermore, our senses are inundated with more information at any one time than we can manage, and so the brain selectively throws away data. We safely presume a reader currently sitting in a chair doesn't feel the pressure exerted by that chair—until we bring that sensation to their attention. It is difficult to catch the brief blackouts that occur when we blink or the blurring that arises when we move our eyes. People experience none of this yet presume their senses provide to them an accurate and complete rendering of the world outside their heads.

The chapter from Yael Granot, Kristyn A. Jones, and Emily Balcetis examines evidence for how our brains tweak perceptual experience.

Defending the idea that perception is not as objective as people often believe it is, the trio of psychologists describe unconscious biases that shape what we "see." They apply the implications of the science for the use of legal evidence in the courtroom: Different participants in a courtroom may literally see the same video evidence differently, thus reaching conflicting verdicts.

Chris Tucker follows with a philosophical counterpoint, asking whether people can in fact rely on their perception. Does the experience that something looks a certain way provide evidence that something really is that way? One prominent epistemological theory answers in the affirmative: We can trust what our perceptual experience tells us, in the absence of counterevidence. Tucker entertains the assertion that unconscious causal influences on perception undermine that theory but argues that the assertion is not convincing. In the next chapter, philosopher Jessie Munton develops something akin to a psychological theory about the ways in which perceptual biases underlie social prejudices. In Munton's view, people perceive and thus judge individuals from other social groups differently due to variations in experience, expertise, and motivation.

The third section of the volume, "Metacognition and Epistemic Evaluation," delves into the assessment of knowledge and expertise in both self and others. David Dunning notes a fundamental problem found in the psychological landscape: People with little expertise often lack the knowledge they need to recognize their own ignorance as well as the true expertise displayed by others. In the following chapter, Nathan Ballantyne explains how that deficit in insight leads to a vexing problem: When non-experts learn that experts do not agree on the right answer to a question, how can they reasonably defer to one side? Since disputes among experts are commonplace, the answer to that question has practical implications.

Hilary Kornblith and Alessandra Tanesini continue the discussion of sources of knowledge worth trusting. In his chapter, Kornblith closely examines an influential philosophical idea about the correct perspective to adopt when evaluating various claims to knowledge. One standard proposal is that such claims need to be assessed from an "objective" perspective: we focus on available evidence or the reliability of the processes at work. It is a cold process based on the facts. Kornblith scrutinizes a radical alternative that assessing knowledge centrally involves something else: feelings of trust and of relationships among people (such as between novices and experts). Tanesini's contribution introduces a new dimension to the collection: the nature of epistemically virtuous and vicious thinkers. Her psychologically

informed philosophical account suggests that virtue and vice are not character traits as much as they are reflected in a person's attitudes—in, for example, the evaluations (i.e., the attitudes) people hold of their own beliefs versus those of others. Tanesini illustrates this by exploring intellectual arrogance and servility.

In the volume's final section, "Cognition and Development," the authors turn to reasoning and knowledge among and about groups. Not only individuals but organizations and nations must often reach useful truths while avoiding costly bias. Hugo Mercier turns to the value of aggregating information across individuals, describing the sources of its benefits as well as why people often fail to see its worth. Deanna Kuhn and Kalypso Iordanou describe fundamental errors people make in reasoning about complex topics. These errors inhibit civil discourse in society, and the psychologists note how education could help, aiding both the individual and the collective. Frank C. Keil and Kristi L. Lockhart explore the question of knowing when something is unknowable by humans and how children come to learn which matters are beyond the grasp of experts.

Interplay Between the Two Disciplines

Although each chapter is a distinct contribution, we find some overarching themes and questions arising from the volume as a whole. These themes and questions reveal something about what epistemology and psychology have to offer one another in conversation and collaboration. We offer two examples we find as readers. We hope and presume that others will see further connections and resonances.

In our first example, epistemology opens up avenues of inquiry for psychologists to pursue. Cheek and Pronin describe perceptions of bias in self versus other, using a definition of bias that comes naturally to psychologists. Devoted to measuring things, psychologists deploy a statistical sense of the term *bias*. They ask whether their measures reveal the true statistical value of a variable or miss the target by a systematic amount. Bias is thus defined in terms of an outcome. In helpful contrast, Kelly and McGrath argue that outcome is secondary and process primary for determining bias. So, to find bias, we instead need to understand the procedure used, rather than whether the outcome varies from the truth. This divergence in definitions of bias raises new questions. Could people, in judging their judgments, share Kelly and

McGrath's intuition and thus assess their bias in terms of the procedure they followed in their reasoning? Or do they instead judge bias in terms of outcome? Do they adopt differing definitions for judging the self versus other?

In our second example, epistemology and psychology cross-pollinate in the opposite direction. Some concepts arise only through empirical psychological research. Take, for example, the notion of naïve realism: people's presumption that they see and understand the world as it really is. In the grip of naïve realism, a person allows that bias exists but mainly if not exclusively in other people. One's own knowledge is much more objective than that of others. Its truth is self-evident. The issue, known from psychological research, is that presumed self-objectivity is often an illusion. People may think they see the world as it truly is while failing to notice how their perceptions and beliefs are based on assumption, interpretation, and bias. Pronin and Cheek consider the phenomenon in the realm of opinions whereas Granot, Jones, and Balcetis show how far naïve realism extends down to visual perception. But what sort of challenges does naïve realism bring for classical issues and arguments in philosophy? How might philosophical topics look different in view of naïve realism? In his chapter, Chris Tucker explores the bearing of empirical findings on one epistemological theory, and reassessment for other themes and theories awaits.

The Co-Editors' Peroration

Reason, Bias, and Inquiry comprises a multidisciplinary meditation for readers who are awash in information but also uncertain about how to manage it all properly. Struggling through the challenges that await our world and society in the coming decades means becoming more clearheaded about being clearheaded. We hope the volume encourages productive discussion and new collaboration.

In sum, gaining insight into reasoning, inquiry, and bias has been an enduring task that each generation throughout history has found to be crucial. The importance of that task has never waned, even as scholars achieve more insight into how reasoning works and how it ought to work. Indeed, in our own generation, making gains in this broad area of investigation may be even more critical than it was for our forebears. After all, the stakes have gone up. Humankind presently fights three horsemen—digital misinformation, political polarization, and a global pandemic—and one wonders if a fourth will

soon arrive on the scene. Perhaps that fourth challenge, one requiring the utmost in impartial and incisive thinking among humans, is here already. It takes the form of climate change.

David Dunning
Nathan Ballantyne

References

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2020). *Cross-country trends in affective polarization* [Working paper 26669]. National Bureau of Economic Research. https://www.nber.org/papers/w26669

Boas, G. (1951/52). Philosophy and Ritual. *Proceedings and Addresses of the American Philosophical Association 25*, 5–17.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146.

Pew Research Center. (2016, June 22). *Partisanship and political animosity in 2016*. https://www.pewresearch.org/politics/2016/06/22/partisanship-and-political-animosity-in-2016/

Pew Research Center. (2019, October 10). *Partisan antipathy: More intense, more personal*. https://pewrsr.ch/3gGRfGp

World Economic Forum (2013). *Global risks 2013*. Geneva, Switzerland: World Economic Forum.

1

Bias

Some Conceptual Geography

Thomas Kelly and Sarah McGrath

Our aim in this chapter is to help illuminate the concept of *bias*.[1] In the first section, we propose that a potentially fruitful project in this area is to identify, from among the many things to which bias is attributed, those that are fundamental in the order of explanation (e.g., given that bias is frequently attributed to both people and their opinions, can one of these be understood in terms of the other?) In pursuing this project, we offer and defend a hypothesis, the *priority of processes*, according to which facts about biased people, groups, and outcomes are typically grounded in more fundamental facts about biased processes and procedures. In the second section, we take up questions about the relationship between a whole's being biased and bias among its parts. In the third section, we offer an account of the concept of bias as it applies to human beings and other entities. According to this account, biases are dispositions to depart from salient symmetry standards in systematic, patterned, or predictable ways. In the final two sections, we offer some remarks about the way in which the concept of bias interacts with the concepts of *reliability* and *knowledge*, respectively.

[1] Recent philosophical work on the topic of bias has primarily focused on the contested phenomenon of implicit bias. (For a sampling of the literature, see Brownstein & Saul [2016]; for an overview, see Brownstein [2017].) The point of departure for this philosophical literature is the explosion of empirical research on implicit bias in recent decades within social psychology (see, e.g., the landmark study Greenwald & Banaji [1995] and, for a popular overview, Greenwald & Banaji [2013]). In contrast, the claims made in what follows are intended to apply to the phenomenon of bias more generally, with implicit bias as a possible special case.

Thomas Kelly and Sarah McGrath, *Bias* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press. © Oxford University Press 2022. DOI: 10.1093/oso/9780197636916.003.0002

What Types of Things Can Be Biased or Unbiased? Are Some of These More Fundamental than Others?

Let's start with a very general question: What types of things can be biased or unbiased?

Clearly, not everything is of the right type. The number 17 has various properties (e.g., the property of being odd as opposed to even). But it doesn't seem to make much sense to claim that the number 17 is *biased* or *unbiased*, and the same seems true of every other number. Similarly, most of the things that natural scientists theorize about, ranging from the very large (e.g., planets and universes) to the very small (e.g., subatomic particles), are not the kinds of things that we would naturally describe as biased or unbiased.

On the other hand, it's striking how many different types of things do routinely get described in this way. It will be helpful to briefly survey some of the more salient categories:

- We predicate bias of people or particular individuals. (We say "He's biased" or "She's biased" with respect to some issue or cluster of issues.)
- We predicate bias of particular individuals in their social roles (e.g., "the biased judge").
- Similarly, we attribute bias to groups or collections of people (e.g., "the biased committee").
- We attribute bias to paradigmatically inanimate objects (e.g., "the biased coin").
- We talk about biased samples or biased data.
- We attribute bias to temporally extended processes, practices, and procedures, as in "a biased admissions process" or "a biased job search."
- We attribute bias to sources of information or putative information (e.g., "MSNBC has a liberal bias," "Fox News has a conservative bias").
- An especially important category for both philosophers and psychologists consists of mental states: We regularly speak of biased perceptions, biased beliefs, biased judgments, biased opinions, and so on.
- Overlapping with the previous category, we frequently attribute bias to the outcomes of deliberative processes; thus, we talk about biased verdicts or biased decisions, and so on.

This list is far from a complete inventory. Nevertheless, it suffices to make the point that many different things can be biased, at least if one takes ordinary thought and talk at face value.[2]

Moreover, notice that even this incomplete list is quite diverse, not only with respect to the range of items that it contains but also with respect to the fundamental categories to which those items belong. For example, on the one hand, we often predicate bias of objects or concrete particulars, as when we predicate it of particular people or coins. But, on the other hand, we are equally happy to predicate bias of things that aren't objects at all. For example, when one says, "The judge arrived at his decision in a biased manner," one attributes bias not to an object or a concrete particular (at least in the first instance) but rather to the process or procedure that the judge used in arriving at the decision.

What should we make of the fact that ordinary thought and talk attribute bias to such a diverse collection of things? Perhaps it's simply a big jumble, and there is not much more to be said on this front beyond that. However, we might try to impose some conceptual order on the jumble by pursuing another possibility. It might be that although many different kinds of things can be biased, some of these are more fundamental than others, in the following sense: When one of the less fundamental things is biased, it has this property in virtue of the relationship that it stands in to something more fundamental, which also has the property of being biased.

For example, suppose that a particular judge is biased with respect to some issue that is presented. Presumably, this is not simply a brute fact about the judge. Perhaps it's like this: The reason the judge counts as biased is that the way in which they arrive at the verdict (or the way in which they would arrive at it) is itself a biased process or procedure. The fact that the judge is biased about a certain question is thus grounded in the more fundamental fact that the way in which they arrive at a verdict is a biased process.

Can this kind of analysis be generalized? We suggest that the following is a potentially fruitful project for philosophers and others to pursue in this

[2] In both ordinary and academic discourse, *bias* and its cognates often—but not always—function like thick terms, in the ethical theorist's sense: For example, to call a judge biased is to make a claim that has descriptive content but one which would not be taken to be evaluatively neutral in any ordinary context. As our list of examples suggests, in what follows we will be particularly concerned with bias in this sense. As the list also suggests, even with this restriction in place the range of things to which people attribute bias is striking. On "thick" concepts, see especially Väyrynen (2017) and Roberts (2013).

area: Given that many different types of things can be biased, are some of these more fundamental in the order of explanation? If so, which?

Here is a comparison for the kind of project we have in mind. Consider the following question:

What types of things can be true or false?

There is an orthodox view about this among philosophers, a view that comes in two parts. According to the first part of orthodoxy, the property of being true is exemplified by many different things, including mental states and cognitive acts (paradigmatically, beliefs and judgments), linguistic entities (paradigmatically, sentences of a natural language), token speech acts (e.g., your asserting, on a particular occasion, that snow is white), and so on.[3] But, according to the second part of philosophical orthodoxy, one of these stands out from the rest as fundamental: In particular, propositions are the fundamental bearers of truth, and anything else that has the property of being true has that property in virtue of the relationship that it stands in to some true proposition. According to this line of thought, to have a true belief is to stand in the believing relation to some proposition that has the property of being true. Similarly, the reason both the English sentence "Snow is white" and the Japanese sentence which is its literal translation count as true is that both sentences are used to express a certain proposition, a proposition that itself has the property of being true. Thus, although the English sentence, the Japanese sentence, and the proposition they express are all true, it is the proposition and its truth that are fundamental in the order of explanation. The fact that the English sentence and the Japanese sentence are true is therefore a derivative matter: They inherit their truth from the true proposition that they express. Perhaps bias exhibits a similar structure.

Next, we offer a hypothesis: At least among the things mentioned thus far, it's typically processes or procedures that are fundamental. That is, although people attribute bias to all of the things mentioned, typically, when something that is not a process is biased, there is some process or potential process in the vicinity that also has or would have the property of being biased; and

[3] Thus, according to philosophical orthodoxy, truth is promiscuous: It doesn't attach to just one thing; it attaches to many. Moreover, notice that, as in the case of things that can be biased, even this incomplete inventory is diverse with respect to ontological category, for it includes states (e.g., beliefs), events (e.g., token utterances), and things that appear to be neither states nor events (e.g., sentences of a natural language).

moreover, the thing that is not a process counts as biased in virtue of the relationship that it stands in to the biased process. It inherits its status as biased from the biased process.

Let's call this hypothesis *The Priority of Processes.*

In order to make the hypothesis more concrete, consider again the biased judge. It seems as though the reason that the judge counts as biased is that the way in which they arrive at a decision (or the way in which they would arrive at a decision) is a biased process or procedure. Our thought that is that when we predicate bias of individuals, it's because we think that they are disposed to arrive at their view, beliefs, judgments, decisions, etc., in a biased manner. A theorist attempting to reverse that order of explanation might make the following claim: "Look, it's not that biased people count as biased in virtue of using biased procedures. Rather, it's that certain people are biased, and the biased procedures are just those procedures—whatever they are—that are used by the biased people." But that seems like the wrong way around.

Consider another class of cases. As noted, it is common to predicate bias of people's judgments, perceptions, decisions, and so on. These judgments, perceptions, and decisions are not themselves processes; rather, they are the outcomes of processes. Here again, our suggestion is that when it's true to describe a judgment, perception, or decision as biased, this is because the process that gave rise to it[4] was biased. According to this line of thought, when someone claims that a judge's verdict is biased, whether this claim is correct depends on whether the process or procedure that the judge employed to reach that verdict itself had the property of being biased. The judge's verdict counts as biased (if it does) because it is the outcome of a biased process or procedure.

Again, someone might try—mistakenly, in our view—to reverse the order of explanation. They might say, "Look, what's fundamental here is the biased verdict. And the process that led to it counts as biased precisely because it delivers a biased verdict, as opposed to an unbiased one." Against this suggestion, here is a straightforward reason for thinking that it is not the biased verdict that is fundamental. Notice that, typically, there won't be anything intrinsic to the content of the verdict that makes it biased: If the same judgment had been reached in an unbiased manner (i.e., by way of an unbiased process), then it wouldn't count as biased. Thus, we can imagine two scenarios.

[4] In some cases, the relevant process might be not the process which originally gave rise to the judgment (perception, decision) but rather the process that sustains it.

In the first scenario, the judge arrives at the verdict that the defendant is guilty because the defendant has an Irish surname, and the judge tends to believe that defendants with Irish surnames are guilty of the crimes of which they have been accused. In this scenario, the verdict that the defendant is guilty counts as a biased verdict. But suppose instead that the judge reaches the same verdict in a different way, by carefully considering all of the available evidence and properly weighing that evidence in order to conclude: The defendant is guilty. If the evidence really does support a guilty verdict, and it is the evidence which is psychologically efficacious for the judge, then the judgment counts as unbiased.

In short, the identical judgment (with respect to its content) might be made by the same individual in two scenarios, but whether it counts as biased or unbiased will differ depending on the process by which it was reached. In this respect, the status of the judgment is derivative or inherited from the status of the process. If that's correct, then it tends to support the hypothesis that processes are more fundamental in the order of explanation (at least compared to things that can be viewed as the outcomes of those processes).

At this point, a clarification concerning the hypothesis is in order. Consider once again the case of the biased judge. In claiming that what is fundamental in the order of explanation is the bias of the process that they employ as opposed to the bias of the judgment that they reach, we intend our claim to be understood as a thesis about the truth conditions of the relevant attributions of bias, as opposed to a thesis about their characteristic epistemology. Typically, observers will not have access to the process that produced the judgment. (Or, at least, they will not have any very direct access to the process.) Rather, what observers will have access to is the outcome of the process, that is, the judgment itself. Because of this, it might be that in the usual case one's evidence for attributing bias includes the token judgment (or a pattern of token judgments); the conclusion that the process being used is a biased process is an inference from what one has to go on, namely the observed judgment or observed pattern of judgments. One sees a pattern of judgments consistently going against a certain kind of defendant, a pattern that one thinks would not have obtained or that would have been extremely unlikely to have obtained if unbiased processes were being used. One then concludes that the best explanation of the observed pattern of judgments is that a biased process is being used to generate those judgments. All of this is to say that the hypothesis that processes are fundamental in this kind of case should be understood in a way that is consistent with the following idea,

which we also think is correct: In the order of discovery (i.e., when it comes to making justified attributions of bias), processes are typically not what is fundamental.

Consider next a challenge to the priority of processes hypothesis[5]:

> Ultimately, one can't make sense of the notion of a biased process without appealing to the notion of a biased outcome produced by that process. Often, when a process counts as biased, this won't be in virtue of its intrinsic properties. Rather, it will be in virtue of its tending to produce biased outcomes or biased patterns of outcomes (perhaps in a modally robust way). That is, even if any particular outcome counts as biased in virtue of being produced by a biased process, the fact that the process counts as biased in the first place depends on the fact that it tends to produce a certain type of outcome or pattern of outcomes, viz. ones that are biased. Therefore, processes aren't fundamental in the order of explanation after all.

This objection raises an important issue. However, we do not believe that it provides a good reason to reject the priority of processes hypothesis, once that hypothesis is correctly understood.

As a first step toward seeing why this is so, notice that the priority of processes hypothesis is perfectly consistent with the following claim:

1. Whether a given process is biased is typically not an intrinsic property of the process: Rather, in typical cases, a biased process counts as biased in virtue of some property of the outcomes (or pattern of outcomes) that it tends to produce.

What the priority of processes hypothesis is not consistent with is the following, stronger claim:

2. Whether a given process is biased is typically not an intrinsic property of the process: Rather, in typical cases, a biased process counts as biased in virtue of the fact that the outcomes (or pattern of outcomes) that it tends to produce have the property of being biased.

[5] Thanks to Alisabeth Ayars here.

We emphasize the distinction between these two claims because we are inclined to think that this is where the truth lies: A given process counts as biased in virtue of its tendency to give rise to certain outcomes, but the key property of those outcomes is not that of being biased. Consider a central case: The relationship between biased samples or biased data, on the one hand, and biased sampling or data-gathering procedures, on the other. What is a biased sample? In some discussions, "biased sample" is simply used as a synonym for "unrepresentative sample." But we take this to be a mistake and not the way that the notion of a biased sample is explicated in the more sophisticated statistics textbooks. (See, e.g., Lane [n.d., pp. 235, 660] for clear discussion of the point.) An unrepresentative sample is one that does not accurately reflect the population from which it is drawn. In principle, a sample might be unrepresentative but not biased: This occurs, for example, when a methodologically impeccable sampling procedure is applied to a population and unluckily generates a sample that does not accurately reflect the population in the relevant respect. (Even if the target population is more or less evenly split between Democrats and Republicans, one ends up calling many more Republicans through simple bad luck.) Rather, a biased sample is one that is generated by a biased sampling procedure—the sample has the property of being biased because it inherits that property from the procedure. A biased sampling procedure is one in which not every member of the target population has an equal chance of being included in the sample. Of course, when one explains why a given procedure or process has the property of being biased, one doesn't need to do that in a way that is completely independent of its propensity to generate a certain kind of outcome. For example, one might characterize a biased procedure as one that tends to produce unrepresentative samples or outcomes. Notice that such a characterization is perfectly consistent with the priority of processes idea, so long as one does not simply identify "unrepresentative sample" and "biased sample."

Consider another central case, that of the coin. Imagine that a fair coin is flipped repeatedly in a fair way. As it happens, the "heads" greatly outnumber the "tails" for a large number of flips. Does that collection of outcomes have the property of being biased, in any theoretically interesting or important sense? We think not. More generally, in many cases, one has some independent grip on the idea of an outcome or collection of outcomes being unrepresentative, where this is not a matter of its being biased in any theoretically interesting or important sense. Whenever we do have some independent grip on an outcome's being unrepresentative, this raises the

possibility of explaining why a given process or procedure counts as biased in virtue of its propensity to produce outcomes like that.[6] And such an explanation will be perfectly consistent with the priority of processes hypothesis.

How far can the priority of processes idea be pushed? We are sympathetic to the idea that processes are at least more fundamental in the order of explanation than people or groups of people and more fundamental than the actual outcomes of processes (whether these outcomes are token judgments, beliefs, decisions, verdicts, etc.). Perhaps there are limits to how far the priority of processes idea can be pushed, and ultimately it will need to be qualified in various ways. But for now, let's leave that as an open question in order to move on and put some further ideas on the table.

Parts and Wholes

A noteworthy feature of our incomplete list of things which can be biased is that some of the items on the list frequently stand in part–whole relations to other things on the list. An obvious example here is people and groups of people: Just as we can evaluate individual judges as biased or unbiased, so too we can evaluate a court that is composed of the individual judges as biased or unbiased (e.g., we might attribute a certain bias to the US Supreme Court). The same holds for processes and procedures, which in a structurally parallel way often stand in part–whole relations to other processes and procedures. The overall admissions process at a university will typically consist of various (sub-)processes, corresponding to the different stages in which the original pool of applicants is progressively narrowed down. Given this, we can evaluate both the overall process as well as the subprocesses with respect to bias.

Consider then cases in which some whole is composed of parts that are themselves the sorts of things that can be biased or unbiased. Often, when the whole is biased, the explanation will be that at least some of its parts (or some critical mass of its parts) have the relevant bias themselves. The fact that a news organization has a certain political bias is because some (or enough) of the individual reporters and editors who work for it have that political bias. Similarly, if the whole is unbiased, the correct explanation for its being unbiased (at least at one level of abstraction) will often be a lack of bias among its

[6] The notion of a symmetry standard, briefly discussed later (see "Biases as . . ."), is one route by which one might attempt to gain some independent grip on the relevant kind of unrepresentativeness.

members, as when the balanced, even-handed coverage of the newspaper is attributable to the same virtues among the members of its staff.

That much seems obvious and familiar. But precisely because it so often works that way, we should be wary of any simple reductionist pictures about the relationship between bias at the level of the whole and bias at the level of the parts. In general, even when one is concerned with wholes that are made up of parts that can be biased, having component parts that are biased is neither a necessary nor a sufficient condition for the whole's being biased. Let's take these points one at a time.

First, in principle, a whole might be unbiased even if its constituent parts are biased to a high degree. Perhaps the most obvious possibility here is when the biases of the parts offset or counteract one another in such a way as to produce a lack of bias at the level of the whole. This could even be due to intentional design—for example, someone could deliberately design an organization or institution to be unbiased in spite of, or even because of, the biases of its constituent parts or members. As a rough comparison, think of the idea behind adversarial systems of justice. In the American legal system, there is no ideal according to which the defense attorney is supposed to be scrupulously neutral between their client and the prosecution; nor is there any ideal to the effect that the prosecution is supposed to be scrupulously neutral. In contrast, many alternative legal systems do not incorporate these partisan elements. It is an empirical question which system does a better job. (For some evidence in favor of adversarial systems, see Thibaut and Walker [1975].) In principle (although, no doubt, this kind of thing is difficult to pull off in practice), a procedure which deliberately incorporates biased parts, even heavily biased parts, might score better when we evaluate the whole. Thus, there can be unbiased wholes that have biased parts.

Conversely, it seems that there can also be biased wholes with unbiased parts.[7] For example, imagine a news organization that specializes in reporting on politics. Presumably, the reporters will have their own personal political opinions, just like everyone else. Suppose that when one considers the reporters as individuals in isolation, and even when one examines them going about their business doing their specific jobs for the organization, one wouldn't attribute bias to them: They do not seem to stand out in any salient way from other people in the same profession whom one would

[7] Perhaps this last point is implicit in recent discussions about institutional racism or various kinds of institutional biases. Even if the individuals who make up an institution are free or relatively free from a certain bias, it doesn't follow that the institution will necessarily lack that bias.

unhesitatingly describe as "unbiased." But suppose further that all or virtually all of the people who work for this particular news organization are like-minded when it comes to politics. They share all of the same political views, or at least, they hold political views that fall within the same relatively narrow band of opinion. (As we are imagining the case, all of these views might even be perfectly reasonable things for them to think, given their evidence, their past experiences and individual life histories, and so on.) Nevertheless, the utter lack of diversity of opinion—a characteristic that only emerges when they are considered not as individuals but as a group—might result, by way of familiar mechanisms, in biased coverage of the news. In this case, it would be correct to describe the news organization as biased, even though one would not attribute bias to the individuals who make it up.

To be clear, we do not believe that this is the usual case. Inasmuch as we think that Fox News and MSNBC have certain biases, we think that those biases are shared, as a matter of sociological fact, by a significant number of people who are responsible for content at those organizations and that the former is due to the latter. However, we also believe that lack of bias at the level of the parts does not guarantee lack of bias at the level of the whole and that it is worth staying alert to the possibility of such holistic bias.

Biases as . . .

Typically, when one attributes bias to an individual or group, one attributes to that individual or group a certain *disposition*.[8] For example, a judge or court that is biased against people of a certain race is disposed to rule against people of that race. To claim that most human beings have *status quo bias* is to attribute to human beings a disposition to favor the status quo merely because it's the status quo. Similarly, to say that a coin is biased in favor of heads is to attribute to the coin a certain disposition or dispositional property.

Like dispositions more generally, biases can be real even during periods when they are not manifested and, indeed, even if they are *never* manifested. (Stock example: It can be true to say that a cup is fragile, even if it is never dropped and so never breaks.) So too with biases: Even if a judge is biased

[8] If one accepts the priority of processes hypothesis, then one will ultimately want to understand this disposition as involving some process(es) that itself has the property of being biased. But we don't wish to presuppose the truth of that hypothesis in this section.

against people of a certain kind, this bias might never manifest itself—if, for example, no one of that kind ever appears in the court or if all of the cases in which they do appear are so clear-cut that the judge's bias is never triggered, in a way that it would be in less clear-cut cases.

Biases are dispositions, but not just any disposition is a bias. Biases are dispositions to depart from normative standards.

Consider the case of a basketball referee. We have the ideal of the calls that the referee ought to make, given what actually occurs in the game that they're officiating and the rules of the sport. Similarly, there is the ideal of the verdicts that the judge or court should reach, given the facts presented in court plus the relevant pieces of law. The calls that ought to be made in the basketball game or the verdicts that ought to be made in court provide a standard which the efforts of actual referees and actual judges might either meet or fail to meet, to varying degrees. The biased referee or judge is disposed to depart from the standard: In a range of possible cases, they will judge in a way that differs from the way in which they should have.[9]

Biases, like most other dispositions, typically admit of degrees. A basketball referee might be egregiously biased against one of the two teams or only slightly and subtly biased against it. In the latter case, they might for the most part officiate the game in the manner of an unbiased referee and be disposed to depart from that standard only in extremely marginal cases in which the actual fact of the matter is difficult to discern for all involved, cases in which they disproportionately favor one of the two teams over the other. Similarly, a coin might be ever so slightly biased in favor of heads or more significantly biased. Any coin which is biased departs from the standard provided by the perfectly unbiased coin, and the extent to which a given coin is biased depends on the extent of the departure from that ideal.

Of course, it's vastly improbable that any actual coin is perfectly unbiased, given sufficiently demanding standards of precision. Assuming that that's the case, does it follow that there really aren't any unbiased coins after all? We think that it would be a mistake to draw that conclusion, for the same reason

[9] Similarly, when we describe a coin as biased, we assume a certain standard, namely, the standard provided by the unbiased coin, which is disposed to land heads exactly 50% of the time (when flipped in the usual way). As this and other examples suggest, we should not assume that the relevant standards are deep features of the true normative order or anything of the sort. A world in which most coins are disposed to land heads half the time need not be better (even in that respect and all else being equal) than a universe in which most coins favor heads over tails. And to the extent that we think that a world containing mostly fair coins is better (say, because of our practice of using fair coins at the outset of American football games in order to determine first possession), we can easily imagine having a different set of purposes that would rationalize the opposite preference.

that it would be a mistake to conclude there really aren't any flat surfaces on the grounds that every surface will turn out to have some bumps on it if examined under a sufficiently powerful microscope. In both cases, close enough is good enough; and in both cases, what counts as "close enough" is plausibly a matter that is both vague and context-sensitive.[10] Similarly—and importantly—we might correctly count some actual people (e.g., judges) as unbiased if they approximate the salient ideal sufficiently closely and even if the way that they fall short of the ideal is exactly the kind of departure that would justify a charge of bias if it were more pronounced than it actually is. Here as elsewhere, one can count as a genuine instance even if one falls short of the Platonic form.

Biases are dispositions to depart from normative standards. But not just any way of being disposed to depart from a normative standard amounts to a bias.

Consider, for example, the unbiased but incompetent basketball referee, who is scrupulously impartial but nevertheless still frequently makes the wrong calls. They thus frequently depart from the relevant standard but not in any systematic or patterned way. (Imagine that they frequently arrive at calls by guessing randomly, so their errors are all over the map: It's not as though their incorrect calls tend to favor one team over the other or players with a certain physical appearance or playing style or anything of the sort.) Notwithstanding their unreliability, it seems as though even arbitrarily large departures from the salient normative standard are consistent with their being unbiased, even perfectly unbiased.

Given that not just any way of being disposed to depart from a normative standard amounts to a bias, we should qualify things further, as follows: A bias is a disposition to depart from a normative standard in a systematic, patterned, or predictable way.

Does this adequately capture the notion of interest? Or do we need to qualify things still further? We believe that the account as it currently stands corresponds to at least one notion of bias that is frequently employed in the social sciences. In fact, given how broadly some social scientists and psychologists use the term *bias*, we suspect that it would be quite difficult, and perhaps impossible, to come up with a less inclusive account (one that

[10] On these points, see especially David Lewis's (1983, Chapter 13) discussion of "flat." Lewis is responding to Unger (1975), who develops and defends a view on which there would be no flat surfaces or unbiased coins.

imposes additional necessary conditions) that does not at the same time exclude at least some of the things that some working scientists would classify as biases. If that's true, then the account as it stands might provide a plausible explication of at least one concept of bias, in something like Carnap's (1950) sense of explication.[11]

Nevertheless, it also seems that there are some dispositions to depart from normative standards in systematic, patterned, or predictable ways that would not ordinarily be counted as biases. Consider, for example, dispositions to systematic mispronunciation. Two of our three children tend to mispronounce certain words or sounds when those words or sounds occur embedded within certain linguistic constructions. Thus, our children are disposed to depart from the norms of correct English pronunciation, in patterned and predictable ways. However, it does not seem correct—at least to our ears—to count the relevant dispositions as biases. Similarly, a person who consistently makes the same kind of mistake when doing long division systematically departs from the norms of arithmetic but is not biased on that account. Thus, it seems as though some dispositions to systematically deviate from salient standards are naturally counted as biases, but others are not. What makes the difference?

Here is a conjecture.[12] Perhaps a bias is a disposition to systematically deviate not just from any salient standard but rather from some salient *symmetry* standard. Consider again the simple paradigm of the biased coin. When the coin is flipped, there are two possible outcomes. It's characteristic of the unbiased coin that there is a symmetry between those two outcomes, inasmuch as they are equally probable. In the case of the biased coin, the symmetry is absent. The same holds when one is concerned with subjective probability as opposed to objective or physical probability. A person who is disposed to invest more credence in the possibility that a fair coin will land heads than in the possibility that it will land tails is biased in favor of heads. A natural thought is that this counts as a bias not because it involves violating a constraint on rational believing—after all, there are many ways of

[11] As understood by Carnap, an explication of a concept does not aim to perfectly capture or accommodate all of the subtleties and nuances of ordinary usage; indeed, it typically will not even be equivalent in its extension to the pre-theoretical notion. Rather, the aim of an explication is to capture the theoretically interesting and important notion in the vicinity, and it makes sense from the standpoint of theorizing to carve things up in this way, even when by carving things up in this way one departs at the margins from ordinary usage.

[12] The subsequent discussion was inspired by a suggestion from Christian List; he should not be held responsible for the way in which we develop it.

doing that that would not count as the manifestation of a bias—but because it involves violating a symmetry constraint on rational believing.

Our conjecture is that the relevant story generalizes. Even in more complicated cases, instances of bias typically involve systematic departures from symmetry standards, while being unbiased involves preserving certain symmetries and invariances. Consider an admissions process. The simplest way to have an unbiased admissions process is when every applicant has an equal chance of being admitted, as when offers of admission are determined by a fair lottery. But, of course, even if some applicants—the better-qualified ones—have a greater chance of being admitted than others, the process might still be unbiased. In that case, even though applicants will differ in their chances of being admitted, other symmetries will be preserved. For example, if an admissions process is unbiased with respect to ethnicity, then applicants of different ethnic backgrounds with equal qualifications will have equal chances of gaining admission: The chance of an applicant's getting in will be invariant with respect to their ethnicity. In the case of the biased admissions process, this symmetry will not be preserved. By contrast, when our son mispronounces certain "r" words, although he is systematically departing from a certain norm, the norm or standard that he is violating does not have the right kind of structure. Our conjecture is this: When a person is disposed to systematically depart from a standard, the more the standard in question is naturally thought of as a symmetry standard, the more natural it will be to consider the disposition a bias.

Again, much more could be said here. But once again, we will move on in order to put some further ideas on the table.

Bias and Reliability

In this section, we address a topic that has already been broached, viz. the connections between being biased and unbiased and being reliable and unreliable. It's natural to think that there are straightforward connections between these notions. After all, claims such as "biased sources of information are unreliable sources of information" or "biased sampling procedures are unreliable" seem like platitudes. However, we think that the connections here are less straightforward than one might initially think, in ways that are philosophically interesting.

We have noted that a person (or a source of information or a process) might be unreliable even if they are unbiased, as in the case of the unbiased but incompetent basketball referee. One common explanation for why someone is unreliable might be that they are biased, but the person might be unreliable for reasons that having nothing to do with bias. In general, unreliability does not entail bias.

What about the other way around? Does being biased entail unreliability? We maintain that the correct answer to this question is not straightforward. On the one hand, we hold that certain generalizations such as "biased thinkers are unreliable" are true, at least when they are suitably interpreted. On the other hand, claims of this sort have to be interpreted so that they are consistent with the following facts. Most cognitive processes are such that whether they are reliable or unreliable depends on the environment in which they are operating. Typically, the same process will be reliable in some possible environments but unreliable in others—a point familiar from discussions of reliabilism in epistemology[13] and from the "rationality wars" in psychology. (See, e.g., Gigerenzer [1991] on the reliability of heuristics in adaptive circumstances.)

Suppose that an individual is biased against a certain group of people. In particular, they are disposed to conclude that people of group F are likely to be G, where G is some negative trait. (We are assuming here that there is no a priori reason to think that people of group F are any more likely to be G; nor does the individual have any significant empirical evidence that they are. It's just a bias that the individual has: When they encounter a person who belongs to group F, they infer that the person is G or that it is likely that the person is G.) Of course, whether that pattern of inference leads the person to mostly true or false beliefs will depend on the environment that they are in. In particular, if they happen to be in an environment in which the people of group F whom they encounter do tend to have feature G, then the biased pattern of reasoning (and the biased person who employs it) will be reliable, at least if we measure reliability in terms of actual relative frequencies of true beliefs among total beliefs. Indeed, the biased person might be more reliable than the unbiased person (who waits for evidence before concluding that this person of group F is G), with respect to the beliefs that they end up with.

[13] For an overview of the issues, see Goldman and Beddor (2016). Perhaps if one gets liberal enough with what counts as a process, one can describe some processes that will be reliable in any possible environment and others that will be unreliable in any environment. But, at best, these will be atypical cases.

More generally, the content of a bias might dovetail with the world in the right way so that the biased thinker ends up reliable.

(Again, the case for thinking that the person is biased, despite their superior reliability, is just this: They didn't have any good reason to think that the Fs would be G. They were just lucky: If they were in an environment in which most of the Fs that they encounter are not G, they would still be disposed to jump to conclusions that the Fs are, and they would end up with many false beliefs as a result of this disposition.)

Does the fact that biased thinkers can end up reliable when their biases dovetail with their environment in the right way mean that we should reject apparent platitudes like "biased thinkers are unreliable"? We don't think so. As indicated, we believe that the apparent platitudes are consistent with the observations just offered, at least when the apparent platitudes are charitably interpreted. There are at least two different (albeit compatible) moves that might be made in order to preserve the intuitive connection between bias and unreliability.

First, we believe that, when properly interpreted, a claim such as "biased thinkers are unreliable" is not a universal generalization. (That is, its truth conditions are not of the form: "For any x, if x is an unbiased thinker, then x is unreliable.") Rather, the better interpretation of such claims is that they are actually generic generalizations, which tolerate exceptions. In this respect, they are similar to generic generalizations such as "Tigers have stripes" or "Dogs have four legs," which are properly counted as true, notwithstanding the existence of tigers without stripes and dogs with fewer than four legs. (On generic generalizations, see especially Leslie [2008]).

Second, one might fall back on the thought that, on the best interpretation of *reliability*, being reliable is not simply a matter of the actual relative frequency of true beliefs among total beliefs. (Again, this is a point that is familiar from the reliabilism literature in epistemology; see, e.g., Goldman & Beddor [2016].) According to this line of thought, on the relevant understanding of reliability, in order to count as reliable, it is not sufficient to actually do well with respect to arriving at true beliefs as opposed to false beliefs. Rather, reliability also includes a certain modal element: In particular, it matters how the biased thinker would fare in certain possible but non-actual circumstances—including, crucially, possible environments in which the content of their bias does not dovetail with the world in the right way.

These then would be two ways of consistently combining the platitudes that posit a connection between bias and unreliability with the observation

that a bias might dovetail with the environment in such a way as to produce a relatively high proportion of true beliefs to total beliefs.

However, it's worth noting why this apparent element of slack between being biased and being unreliable arises in the first place. Generally speaking, the extent to which a given person or cognitive process is reliable will not be an intrinsic feature of the person or process. Rather, it will be a relational feature of the person or process, one that is tied to the environment that the person is in or the environment in which the process operates. Consider the careful, responsible thinker who scrupulously attends to the evidence in arriving at their beliefs. In our world, these cognitive habits and dispositions effectively promote the aim of having true rather than false beliefs. (Or, at least, they fare better with respect to having true rather than false beliefs than those who ignore their evidence and arrive at their beliefs in a cursory or haphazard manner.) However, when one considers how the same thinker fares in a possible world run by a Cartesian evil demon bent on deceiving its inhabitants, one realizes that these same habits and dispositions are completely unreliable—in fact, they lead one to do worse than those who form their beliefs in a haphazard manner, precisely because the demon has seen to it that the evidence available in the world is systematically misleading as to the world's true character. The lesson: In general, it seems as though whether a thinker or a way of arriving at beliefs is reliable, or the extent to which they are reliable, is a relational matter. It depends not just on the thinker or process itself but on the world.

By contrast, on what seems like the most natural way of thinking about biases, whether someone has a given bias is not a relational matter, in anything like the same way. If we transport a person to another possible world but hold their psychology fixed, then they will have all the same biases as before (even if the world is very different). Being biased or unbiased, having or lacking a certain bias—at least at one level of abstraction, these seem like intrinsic features of the person. In contrast, the effects of having or lacking a certain bias on the person's cognitive projects will often vary significantly with the world or environment that they are in.[14]

[14] On a traditional view about the metaphysics of dispositions, dispositions are intrinsic properties of their possessors. Given this traditional view and given the thesis that biases are dispositions, the claims that biases are intrinsic features of a person who has them and that merely transporting a person from one environment to another without altering their intrinsic properties will leave their biases intact follow immediately. For classic statements of the idea that dispositions are intrinsic, see especially Johnston (1992, 232–234) and Lewis (1997, 147–148). McKitrick (2003) is a challenge to the idea.

Bias and Knowledge

We will conclude with some remarks about the connections between bias and knowledge.

A natural thought is that being biased—at least, when the bias in question is sufficiently strong—excludes knowing. Imagine a judge whose bias against a defendant is so strong that they would believe that the defendant is guilty regardless of what the evidence suggests. When weak evidence of the defendant's guilt is presented in court, the judge concludes that the defendant is guilty, and their drawing that conclusion is a manifestation of their bias. This seems to guarantee that the judge's belief is not knowledge: Even if the defendant actually is guilty, and so the judge's belief is true, it still fails to count as an instance of knowing.

While we agree that this is the right thing to say about the case as described, we also think that caution is in order when it comes to drawing general lessons. One common way in which cognitive biases might manifest themselves is by making our beliefs insensitive to the truth, in the technical sense of *insensitive* employed by epistemologists. In this sense, your belief that p is insensitive if and only if the following condition holds: If p had been false, you would still have believed p. The judge's belief that the defendant is guilty is insensitive because even if the defendant had been innocent, the judge would still have held the belief that the defendant is guilty. In this way, the judge's belief fails to track the truth because of bias.

It is tempting, then, to conclude that bias excludes knowledge in virtue of making our beliefs insensitive to the truth. Although this would be a theoretically satisfying story about the relationship between bias and knowledge, we believe that it is too simple, for there are compelling reasons to think that sensitivity is not a necessary condition for knowing: Even if one of your beliefs is insensitive and you would hold it regardless of its truth, it might still amount to genuine knowledge.[15]

Consider cases of asymmetric overdetermination. A parent watches a young child playing normally; the parent can plainly see (and, thus, knows) that the child is alive and well, just as anyone else who is viewing the same scene can know the same proposition. Perhaps, however, the parent's belief

[15] The idea that sensitivity is a necessary condition for knowledge is central to Nozick's (1981) "tracking" account of knowledge. For criticism of the idea, see especially Sosa (1999), Williamson (2000), and Kripke (2011, Chapter 7), among others.

that the child is alive and well is insensitive: If the child were not alive and well, the parent would still believe this because the parent is so deeply invested in its being true that the child is doing well. (If credible evidence began to emerge that the child was not alive and well, this would trigger psychological mechanisms that would lead the parent to dismiss that evidence or explain it away so as to allow for the retention of the belief. As we are imagining the example then, it involves an extreme case of bias.) However, even given this stipulation, it seems that, intuitively, the parent can know that the child is alive and well on the basis of straightforward and unproblematic observation, as things actually stand.[16]

Moral: Even if a bias is sufficiently strong to make a given belief inevitable, it does not follow that that belief is not knowledge. Why does this matter?

Here is one reason. Philosophers sometimes propose methodological norms that are clearly intended to safeguard inquiry from being corrupted by various kinds of biases. Consider, for example, the method of reflective equilibrium, which for decades has been the most popular account of how moral inquiry should proceed among moral philosophers.[17] As characterized by Rawls and his followers,[18] the correct starting point for moral inquiry consists of the totality of our considered judgments about morality, in a semi-technical sense of "considered judgment." In this semi-technical sense, a judgment counts as "considered" only if the person making the judgment does not stand to gain or lose depending upon how the question is answered (see, e.g., Rawls, 1971, p. 48; Scanlon 2002, p. 143). Such conditions are clearly intended to safeguard inquiry from the potentially distorting effects of bias. On this view, you should set aside or bracket any moral judgment that aligns with your self-interest or when you are invested in the relevant question being answered in one way rather than another.

However, we think that this is bad advice. Among other things, notice that it conflicts with the following methodological norm, which we take to be true:

[16] Significantly, Nozick (1981) himself would admit that the parent can know in these circumstances; see especially pp. 179–185, where the sensitivity condition is relativized to methods of belief formation in order to accommodate the intuitive verdicts about structurally similar cases.

[17] We date its ascendancy from the publication of Rawls (1971), which was enormously influential in this as well as in other respects.

[18] In addition to Rawls (1971), see especially the characterizations offered in Rawls (1975/2001, Chapter 15), Daniels (1996, 2018), and Scanlon (2002, 2014). For critical discussion, see Kelly and McGrath (2010).

> **The Knowledge Platitude**: If you know something that is relevant to a question that you are trying to answer, then you should take that information into account in arriving at a view.

In any case in which you know something that aligns with your self-interest, the knowledge platitude will instruct you to take that piece of information into account, even though the relevant judgment does not satisfy the conditions for being a considered judgment, in the semi-technical sense.[19] Precisely because the relationship between bias and knowledge is not as straightforward as one might have initially thought, norms of the sort endorsed by Rawls and others threaten to exclude too much.

Let's briefly see how this difference plays out in the context of a concrete example. Consider the following proposition:

> A person of color should not receive lesser consideration in virtue of being a person of color.

Notice that, for a person of color, this judgment is heavily bound up with their own interests.[20] On the face of it, it seems like this judgment will fail to qualify as a considered judgment for a person of color for that reason and thus will be one that they are required to bracket or set aside when it comes to, for example, thinking about which moral theories they should accept. We think that that's the wrong result, however. On the contrary, we believe that it might be perfectly reasonable for a person of color to take this judgment into account in attempting to figure out which moral theories they should accept (e.g., by taking it to count in favor of theories that entail that it is true as opposed to theories that entail that it is false). We think that the reason for this is as follows: notwithstanding the fact that it will typically be very much in the self-interest of a person of color that this proposition is true as opposed to false, they will typically have a high degree of justification for the belief that it is true; indeed, the fact that the proposition aligns with their self-interest is perfectly compatible with their knowing that it's true. And if they know that it's true, then, we submit, not only is it rationally permissible for them to take

[19] For fuller development of this criticism, including a detailed consideration of particular examples, see McGrath (2019, Chapter 2). For considerations in favor of the knowledge platitude, see especially Williamson (2000) and Kelly (2008).

[20] Of course, we do not mean to suggest that this proposition is not bound up with the self-interests of others as well.

it into account in their deliberations, but it would be a methodological mistake for them to set it aside or bracket it.

* * *

To be sure, none of the considerations offered in this section cast doubt on the compelling idea that our biases often constrain what we are in a position to know. What they do suggest is that a subtler account of the way in which the concept of bias interacts with the concept of knowledge will be needed in order to arrive at more adequate methodological norms, as well as for other purposes.

Conclusion

Especially compared to the amount of attention that bias has received from psychologists in recent decades, the topic has been relatively underexplored by philosophers. Recently, this has begun to change, although it is fair to say, we think, that the range of questions in this area that philosophers have pursued has been somewhat limited. In addition to the specific proposals offered here, our hope is that one contribution of the present chapter is to draw attention to the range of broadly conceptual issues about bias that philosophers are well suited (even if not uniquely well suited) to address (e.g., questions about possible relations of explanatory priority among the diverse types of things that can be biased or questions about the structural connections that obtain or fail to obtain between bias and knowledge). Throughout our discussion, we have emphasized just how much our specific proposals leave open. Thus, even if these proposals are on the right track, much remains to be done—and all the more so if they are not.

Acknowledgments

Ancestors of this chapter were delivered as talks at Rutgers, Fordham, Union College, and a Brown-Princeton workshop held at Princeton University. We are grateful to the audiences present on those occasions for their feedback; special thanks to Peter Graham and Katia Vavova for serving as commentators at Rutgers and at Princeton, respectively. Finally, thanks to Robert Audi, Alisabeth Ayars, Nathan Ballantyne, Grace Helton, Mark

Johnston, and an anonymous reader for written comments and/or helpful conversations on earlier versions.

References

Brownstein, M. (2017). Implicit bias. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2017 ed.). Stanford University. https://plato.stanford.edu/archives/spr2017/entries/implicit-bias/

Brownstein, M., & Saul, J. (Eds.). (2016). *Implicit bias and philosophy* (2 vols.). Oxford University Press.

Carnap, R. (1950). *Logical foundations of probability*. University of Chicago Press.

Daniels, N. (1996). *Justice and justification: Reflective equilibrium in theory and practice*. Cambridge University Press.

Daniels, N. (2018). Reflective equilibrium. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2018 ed.). Stanford University. https://plato.stanford.edu/archives/spr2018/entries/reflective-equilibrium/

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." *European Review of Social Psychology*, *2*, 83–115.

Goldman, A., & Beddor, B. (2016). Reliabilist epistemology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 ed.). Stanford University. https://plato.stanford.edu/archives/win2016/entries/reliabilism/

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27.

Greenwald, A. G., & Banaji, M. R. (2013). *Blindspot: Hidden biases of good people*. Delacorte Press.

Johnston, M. (1992). How to speak of the colors. *Philosophical Studies*, *68*(3), 221–263.

Kelly, T. (2008). Common sense as evidence: Against revisionary ontology and skepticism. *Midwest Studies in Philosophy*, *32*(1), 53–78.

Kelly, T., & McGrath, S. (2010). Is reflective equilibrium enough? *Philosophical Perspectives*, *24*, 325–359.

Kripke, S. (2011). *Philosophical troubles: Collected papers* (Vol. 1). Oxford University Press.

Lane, D. M. (n.d.). Introduction to statistics. In *Online statistics education: An interactive multimedia course of study*. http://onlinestatbook.com/

Leslie, S.-J. (2008). Generics: Cognition and acquisition. *Philosophical Review*, *117*(1), 1–47.

Lewis, D. (1983). *Philosophical papers* (Vol. 1). Oxford University Press.

Lewis, D. (1997). Finkish dispositions. *The Philosophical Quarterly*, *47*(187), 143–158.

McGrath, S. (2019). *Moral knowledge*. Oxford University Press.

McKitrick, J. (2003). A case for extrinsic dispositions. *Australasian Journal of Philosophy*, *81*(2), 155–174.

Nozick, R. (1981). *Philosophical explanations*. Harvard University Press.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Rawls, J. (2001). *Collected papers*. Harvard University Press. (Original work published 1975)

Roberts, D. (2013). Thick concepts. *Philosophy Compass*, *8*(8), 677–688.

Scanlon, T. (2002). Rawls on justification. In S. Freeman (Ed.), *The Cambridge companion to Rawls* (pp. 139–167). Cambridge University Press.

Scanlon, T. (2014). *Being realistic about reasons*. Oxford University Press.

Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives*, *13*, 141–153.

Thibaut, J., & Walker, L. (1975). *Procedural justice: A psychological analysis*. Lawrence Erlbaum Associates.

Unger, P. (1975). *Ignorance: A case for skepticism*. Oxford University Press.

Väyrynen, P. (2017). Thick ethical concepts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2017 ed.). Stanford University. https://plato.stanford.edu/archives/fall2017/entries/thick-ethical-concepts/

Williamson, T. (2000). *Knowledge and its limits*. Oxford University Press.

3

Yo-yo Rationality Attributions

Roy Sorensen

As a philosopher, I have colleagues who delight in demonstrating that I am inconsistent—regardless of my thesis. When my mother learned of this, she offered advice based on years of experience managing counter-suggestible children.

MOTHER: Try reverse psychology.
ME: Adults would never fall for such a simple-minded trick!
MOTHER: Sorry dear, I did not realize philosophers are so smart. Do not try reverse psychology.
ME: Then I will try it!

My plan is to confess my worst inconsistency and then have you prove that I am consistent.

Reciprocal Self-Defeat

Confession: I am inconsistent about, of all things, rationality. Worse, my inconsistency is directional. When interpreting people, I maximize their rationality because rationality is a precondition for explaining and predicting actions on the basis of beliefs and desires. Up with rationality!

Yet I minimize rationality because its ingredients are expensive. Rationality depends on attention, self-control, computation, and memory. Since each is in short supply, I postulate only as much rationality as needed to explain actions. Down with rationality!

Worse yet, my opposed conceptions of rationality imply each other. "Rationality is a scarce resource" invokes cost—a classic economic concept that is analyzed with the preferences of rationally self-interested agents. If

Roy Sorensen, *Yo-yo Rationality Attributions* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press. © Oxford University Press 2022. DOI: 10.1093/oso/9780197636916.003.0004

rationality is a cost, then there will be a rational agent making decisions that are less than fully rational.

Most economists refuse to compromise on rationality (Becker, 1976, Chapter 1, pp. 11–13). Rationality is a framework for pinpointing action along the coordinates of belief and desire. To work locally, rationality must be global.

The framework is compatible with ignorance. Research takes time, and time is money. So the notorious ignorance of voters is predicted rather than precluded by rational expectations theory (Downs, 1957).

In *The Myth of the Rational Voter*, Bryan Caplan criticizes colleagues who limit their dumbing down of voters to rational ignorance. We voters do not merely avoid the cost of learning truth. We "*turn off* our rational faculties on topics where we don't care about the truth" (Caplan, 2007, p. 2). We actively indulge in the excitement and consolations afforded by irrational worldviews (Akerlof, 1989). The pleasures of ideological fantasies are safely indulged in the consequence-free environment of elections. My vote only has a negligible chance of affecting the outcome. And even if it does, the price of my irrationality is paid by a large group rather than by me alone.

Caplan thinks the voter's irrationality is limited to circumstances in which they cannot hurt himself. At some level, the voter is always patrolling for signs of danger. This standby rationality intervenes when the stakes are raised.

The basis of the intervention is elusive. Caplan admits that the basis cannot be the agent's antecedent beliefs about self-interest, for these beliefs are frequently affected by the irrational worldview. For instance, religions commonly portray agents as immortal. Secular concerns are trivial in comparison to the afterlife. It is prudent for the Islamic jihadist to die in battle and imprudent for him to surrender. According to Caplan (2007, p. 128), the prospect of certain death causes the jihadist to change his belief about what is in his self-interest. But why does the prospect of death cause conversion to a secular conception of self-interest rather than a reinforcement of the religious conception? If natural selection wires us to lose faith when put to the test, then the secular estimate of self-interest is already entrenched.

Caplan allows irrationality at the level of voting because he thinks it can be restricted to harmless domains. But the empirical evidence of irrationality applies almost as well to personal health and finances. Once our dam against irrationality is cracked, the puddle will not stay small.

Grammatically, *rational* is an absolute term (Sorensen, 1991). Just as a flat surface is free of bumps, curves, or other irregularities, a rational agent is

free of inconsistency, circularity, and other irrationalities. So two rational agents can no more differ in their degree of rationality than can two flat surfaces differ in their degree flatness. Only irrational agents can differ in their degrees of rationality.

Rationality is a feature of the framework for minimizing cost, not a factor within that framework. The economist does not treat cost as a mere cause. Cost is a reason. Austerity with rationality is a budgetary incoherence. "Rationality is a cost" entails "Rationality is a prerequisite for attributing beliefs and desires."

Little wonder that attributions of irrationality stimulate accusations of self-defeat. If I think my audience is irrational, what is the point of arguing with them? Worse, if I think I am irrational, why should they believe what I am saying?

Alas, self-defeat also threatens the view that rationality is a prerequisite for attributing beliefs and desires. Maximizing rationality requires maximizing the rationality of one's "adversaries"—who will turn out agreeing on a priori matters. There would be no market for economics teachers, advisors, and theoreticians. By arguing that rationality is a prerequisite, I take for granted that my adversary disagrees with me about a corollary of the maximizing view: "All agents are rational." This "principle of charity" is an a priori principle of interpretation. Rejecting any a priori principle is irrational because one possesses all the evidence needed to avoid the error. The mistake is internal. If my adversary had been more rational, then he would have understood that everyone must be perfectly rational! . . . Oops! (Sorensen, 2004).

Economists who advocate maximizing rationality inadvertently reveal a belief that rationality is a scarce resource (Caplan, 2007, p. 13). When the American economist Paul Samuelson wrote as a theoretician, he took rationality to be a precondition of explaining consumer choice. When those consumers were his students, Professor Samuelson assumed rationality was in limited supply. He struggled to correct their systematic biases, inattention, and fallacious thinking. "Rationality is a prerequisite" and "Rationality is a cost" form an unstable mix in Samuelson's volatile boast: "I don't care who writes a nation's laws—or crafts its advanced treaties—if I can write its economics textbooks" (Skousen, 1997, p. 150).

In sum, I subscribe to a pair of principles about rationality that are jointly inconsistent: "Rationality is a precondition for explaining action in terms of

belief and desire" and "Rationality is a scarce resource." These two principles are also individually self-defeating because they imply each other.

This mutual dependence prevents a test of strength. In the case of independent antagonisms, conflicting forces can be pitted against each other. For instance, a curious motorist can test whether an accelerator is more powerful than a brake by simultaneously flooring both pedals. Since his car does not move, the brake is proven more powerful than the accelerator. But my conflicting beliefs about rationality rely upon each other. The situation is more akin to braking by downshifting. What makes the car slow makes the car go.

Nevertheless, there might be an a priori test of relative strength. One candidate measurement exploits an asymmetry in the nature of the mutual reliance between "Rationality is a prerequisite" and "Rationality is a cost." "Rationality is a cost" semantically relies upon "Rationality is a prerequisite for attributing beliefs and desires." But the prerequisite principle only pragmatically depends on the cost principle. A maximizer of rationality could consistently cling to the prerequisite principle by abstaining from debate about whether agents are rational. Admittedly the silenced maximizer would be unable to acknowledge that people disagree about whether rationality is a prerequisite for understanding action in terms of belief and desire. The maximizer would also have trouble explaining why students pay tuition to learn about principles of rational choice. The maximizer will also have trouble understanding his past. After all, he appears to recall being enlightened by "Rationality is a prerequisite for understanding action in terms of belief and desire." So his educational memory seems delusive. Yet the maximizer would have the solace of internal consistency.

In contrast, even extremism cannot rescue "Rationality is a cost." No empirical investigation can confirm a hypothesis that presupposes a proposition incompatible with it. So unlike "Rationality is a prerequisite for belief and desire," "Rationality is a cost" has an internal inconsistency that threatens to preclude any empirical support.

A girl trying to walk out of a forest has a consistent directional desire: Travel in a straight line. Nevertheless, she walks in a big circle. With each step, she slightly undoes what was done, returning to her starting point. She did not notice a bias in her stride. I lack this excuse. My directional inconsistency about rationality is a big a priori mistake.

How big? As big as that of a driver surprised by a yellow traffic light. He alternates between the accelerator and the brake. This rapid cycle of

self-cancellations is a sign that the driver lacks a complete strategy for balancing the risk of a T-bone collision (incurred by speeding through a red light) and a rear-end collision (incurred by abrupt braking). He flits between opposite tactics, undoing what was done, redoing what was undone.

At least the dithering driver has the excuse of not being forewarned. And he quickly settles down to one tactic or the other. With me, advance notice does not stop the ambivalence. And I never settle down. For instance, I know that an encounter with a "weak-willed" alcoholic will precipitate yo-yo rationality attributions. With my left foot, I apply the brake against attributing rationality. After all, the alcoholic says he acts contrary to his best interest. But there he is, sterilizing his self-incriminations with sips of alcohol. He chooses vodka because it is cheap and potent and looks like water. His efficiency at inebriation prompts the adage, "Actions speak louder than words." With my right foot, I floor the accelerator and maximize rationality. In my rearview mirror I see the rough edges of the alcoholic recede, leaving a perfect dot of rationality in the distance.

My dissonance has been diagnosed in six ways. First, the rational expectations theorist says I suffer from faint-heartedness. I should reject "Rationality is a cost" and become an unrestrained maximizer of rationality. Second, others say the inconsistency is a linguistic illusion. They say "Rationality is a prerequisite for attributing belief and desire" is a tautology such as "An edge is a prerequisite for checkmating a lone white king with only a black king and queen." A tautology is made true by virtue of the meanings of its words and so is compatible with every consistent proposition. A third diagnosis is that the contradiction is genuine but functional. The pain of inconsistency keeps me exploring, enjoying the benefits of both perspectives. A fourth diagnosis is that I am old-fashioned. I should update my scientific views and recognize that rationality is as obsolete as phlogiston. Just as chemists showed there was no phlogiston, neuroscientists have shown there is no rationality (Churchland, 1986). The fifth diagnosis, instrumentalism, softens this bad news for common sense by rebranding rationality as a tool that can be deployed without metaphysical commitment to the existence of rational agents (Dennett, 1987). The sixth diagnosis combines themes. My ambivalence is characterized as cycles of idealization and deidealization.

Not all of these assessments can be right. But they all illuminate the question by articulating possible answers and challenges to its presuppositions.

Proposal 1: Maximize Rationality Without Restraint!

If I kept my foot on the accelerator, I could be as consistent as Socrates (or at least the Socrates portrayed in Plato's *Protagoras*, 351a–358d). He denies that there is weakness of will. People always act on their strongest desire. This is what they want, all things considered.

If people are always doing what they most want to do, reasons the egoist, then each person pursues their own self-interest. Psychological egoism, like theism, is a doctrine familiar to freshmen prior to their first philosophy class. They already know how to interpret action as the maximization of perceived self-interest.

Socrates foresaw freshman foreknowledge. According to his doctrine of recollection introduced in the *Meno*, Socrates merely draws out pre-existent knowledge (as his mother, a midwife, drew out pre-existent babies from pregnant women).

Nevertheless, the same students who are precocious with psychological egoism are equally imaginative with hypothetical counterexamples. Journalists supplement this corpus of possible altruists with actual specimens ("Newark Mayor Cory Booker Taken to Hospital," 2012).[1] Students polish these stumbling blocks for egoism into stepping stones toward altruism.

But students are restless. Instead of standing on their shining exemplars of altruism, they probe for cracks in their counterexamples to psychological egoism. Students who earlier dissolved the problem of the willed alcoholic with the principle that an agent must act on his dominant desire realize that this solvent can be applied to paradigm cases of altruism. Just posit a vicarious pleasure: The "altruist" most wants to please others. The resolute psychological egoist concludes there is only a difference in taste between the "self-sacrificing" altruist and the self-absorbed alcoholic.

Suppose you resolve to become living testimony against psychological egoism. Your dominant desire is to become a counterexample. If you succeed, you will do what you most want to do. But then you have confirmed rather than disconfirmed psychological egoism. His doctrine is irrefutable!

[1] Some students go on to conduct experiments to prove that human infants are intrinsically motivated to help others (Tomassello, 2019, pp. 220–230).

Proposal 2: Co-opt Maximizing as a Tautology!

This sophisticated defense of psychological egoism is often deflated as a retreat into a tautology. The egoist began with a bold empirical thesis and, under the threat of counterexamples, quietly changed the meaning of key terms so that psychological egoism becomes true by (re)definition.

This suggests a linguistic strategy for resolving the dilemma between maximizing rationality and minimizing rationality: Show that "Rationality is a prerequisite for belief–desire attribution" is a tautology. A tautology cannot conflict with any consistent proposition. So the discovery that "Rationality is a cost" entails "Rationality is a prerequisite for belief–desire attribution" turns out to be harmless.

Tautologies can be used to counsel acquiescence to evidence. "Bishops only move diagonally" can be used to show that someone disturbed the chess pieces on a board that now sports two white bishops on white squares (with no promoted pawns). A tautology can have the generality of empiricism: "Facts are facts," says the experimentalist who has slain a beautiful theory. Tautologies can also exhort psychological search. "All agents are rational" cheers on the quest for a reason behind apparently crazy behavior. When the economist sees his aunt flail near a basement window, he does not give up the search for reasons: "My aunt's apparently aimless motions must be directed toward some purpose. Aha! She was clearing away spider webs that were invisible from a distance."

Any substantive implications of uttering a tautology comes from the fact that the tautology was uttered, not the content of that empty utterance:

WIFE: Why have only my daughters made their bed?
HUSBAND: Boys are boys, and girls are girls.
WIFE: [silence]
HUSBAND: Well, boys are BOYS, and girls are GIRLS.
WIFE: [SILENCE]
HUSBAND: Boys, make your beds!

The husband's initial response invites his wife to explain the difference in bed-making with a stereotype about gender: "Girls are tidy, and boys are messy." His wife spurns the invitation by passing on her turn to speak. Silence is here a rebuke by omission. Conversation is a journey with an expected rate of progress. When your partner stops in her tracks, the effect is negative

rather than neutral. From the perspective of a stationary bystander, absence of saying is saying nothing. Within the moving party, however, the wife's refusal to move forward is a challenge to her husband's unspoken empirical principle. Her husband could escalate his conversational implicature to the outright assertion of "Girls are tidy, and boys are messy." Instead, he diplomatically underscores the tautological status of "Boys are boys, and girls are girls." To the bystander, emphatic repetition of a tautology is marching in place. But on the moving walkway of conversational expectation, standing in place is effortful retreat. His wife is not fully mollified. She continues her boycott on explicit participation in the conversation. Her husband could now take offense at her unwillingness to let him exit gracefully. Instead, he capitulates to this follow-up dose of silence. The couple has engaged in a substantive dialogue despite the absence of substantive assertions. Tautologies speak only in the way silence speaks—between the lines.

Tautological generalizations seem philosophical because they express general approaches and attitudes. Resolution is expressed by "A promise is a promise." Fatalism is conveyed by uttering "Whatever will be, will be." The tautology "There is a reason for everything" can express determination in one conversation and resignation in another.

The rationally mandatory reaction to a tautology is incorporation. If the creationist really thinks that "Only the fittest survive" is a tautology, then he is an evolutionist. It may be misleading advertising for him to second the evolutionist's slogan. But "Only the fittest survive" is, by his lights, the literal truth. The same applies to pessimists about rationality who think "All agents are rational" is a tautology. They think rationality is maximized by being minimized (even if saying so is paradoxical phrasing—as in the case of a judge who echoes the prosecutor's "Murder is murder" and then acquits the defendant on the grounds that the homicide was justified self-defense).

For some tasks, maximizing is entailed by minimizing. Disassemble a rectangular chocolate into its basic components with the minimum number of breaks (using only horizontal or vertical snaps, never cutting through a square). If you succeed in minimizing the number of breaks, you will, thereby, maximize the number of breaks.[2] Just as minimizing breakage entails maximizing breakage, minimizing rationality entails maximizing rationality.

[2] The rectangular chocolate bar has $m \times n$ squares. With each break, whether horizontal or vertical, you increase the number of pieces by one. Since you began with one piece, you inevitably finish in $(m \times n) - 1$ steps. Hidden constants are the foundation stones of magic tricks.

Is the maximizer silenced by having his position exposed as a tautology? Is "All agents are rational" just a philosophical platitude?

Well, one man's tautology is another man's mathematics. According to the maximizer of rationality, every agent's actions must plot along the axes of beliefs and desires—just as every figure must fit into the Cartesian coordinate system. The right attitude toward "Rationality is a prerequisite for attributing beliefs and desires" is the same attitude we have toward the equation for the Euclidean distance between two points of the plane. An environment may be too unstable to make the principle useful. But no state of affairs can count as a counterexample.

Through applied mathematics, we get an a priori discipline contributing to a posteriori predictions. Economists make surprising predictions about how the alcoholic *Homo economicus* will respond to taxes on alcohol. So the economists are not nursing a trivial tautology. They have swaggered into history and sociology, applying "Rationality is a prerequisite for attributing beliefs and desires" unblinkingly to crime, pollution, marriage, and the arms race.

How is this possible? A tautologous premise can always be deleted from an argument without affecting its validity.

Although tautological premises are never needed to deduce a conclusion, inference rules are always needed to make the transition from premises to a conclusion. Rules are neither true nor false. So they cannot contradict any facts of psychology.

When writing his dissertation, Paul Samuelson was embarrassed by the tendency of economic principles to devolve into tautologies. He tried to interpret them as laws of nature, akin to those in physics. He recalls that Stanislaw Ulam

> used to tease me by saying, "Name one proposition in all of the social sciences which is both true and non-trivial." This was a test that I always failed. But now, some thirty years later, on the staircase so to speak, an appropriate answer occurs to me. The Ricardian theory of comparative advantage; the demonstration that trade is mutually profitable even when one country is absolutely more—or less—productive in terms of every commodity. That it is logically true need not be argued before a mathematician; that it is not trivial is attested by the thousands of important and intelligent men who have never been able to grasp the doctrine for themselves or to believe it after it was explained to them. (1969, p. 3)

Since tautologous premises contain no information, the informativeness of economics comes from rules about how to handle information. A positive rule, such as modus ponens, licenses an inference. A negative rule forbids it. Ricardo's principle of comparative advantage is a caution against the fallacy of composition (inferring the whole shares the properties of its parts). Consider a pair of sailors marooned on an otherwise uninhabited island. The young sailor can do everything better than the old sailor. We are tempted to infer that there can be no profitable commerce between the young man and the old man. But the old man's universally inferior labor is still a reservoir of time. The old man can free the young man to engage in more useful pursuits. Although the young man is a somewhat better gardener, he is a far better hunter. If the old man does the gardening that would have consumed the young man's time, the young man can instead do more hunting.

But wait! The connection between tautologies and inference rules flows both ways. Attributing irrationality to one's inferences is no more charitable than attributing irrationality to one's premises. And if people are irrational enough to reject Ricardo's principle of comparative advantage, why are they rational enough to have their decisions predicted by their conformance to it?

This inconsistency leads many academics to scoff at economics. But when they invest their retirement funds, they put their money where their mouths ought to be. Academics funnel a fortune into index funds that apply Paul Samuelson's efficient market hypothesis. His random walk theory of stock market prices is epitomized by a joke. A merchant and an economist are walking down a street. The merchant exclaims, "Look, there is a 20-dollar bill on the pavement!" The economist chides the merchant, "No, there could not be. Someone else would have picked it up." According to Samuelson, public clues about the price of future stock prices are like 20-dollar bills lying on the street. Such clues would instantly bid up the price of the stock. Consequently, the actual price of the stock is the best measure of its value (since it reflects all the available information, not just the information available to particular individuals). The practical consequence is that one should not try to out-predict the market about the price of stocks. Investors should instead buy at random, thereby acquiring a diversified portfolio of stocks, as a long-term investment. This would minimize transaction costs while retaining the stock market's lucrative average return. The stock market is a casino in which the dice are lightly loaded in favor of the customer.

Proposal 3: Regard Yo-yo Attributions of Rationality as Functional!

I like the efficient market hypothesis, but I do not like how I defend it. Under cross-examination from my wife, I do a dialectical dance. To exhibit the random walk, I leap to "Rationality is a prerequisite for attributing beliefs and desires." When pressed on whether I think investors are rational, I dip to the cost view, claiming that the investors are rational when the stakes are high enough to warrant the research costs. When pressed on how lazy thinkers could recognize when the stakes are high enough, I spring to a higher-order rationality that monitors whether circumstances warrant putting on your thinking cap. My wife wonders whether a frugal explainer can help himself to the luxury of back-up rationality. If I am so rational as to be able to discern when to be rational, then my second-order rationality collapses into first-order rationality. I was rational all along, negating any savings in computation, attention, and memory. Alarmed by the instability of my top-heavy tower of rationality, I backpedal to the claim that the market is rational as long as a small percentage of investors are rational (for the logical minority are motivated to quickly exploit, and thereby correct, errors by the majority). When pressed on whether I think this astute minority is always rational, I further backpedal to the claim that membership in the elite shifts from investment to investment. I cannot keep my story straight! I talk from both sides of my mouth. But I must mean it, for I am one of the investors, putting my money where my mouth is.

The behavioral economist consoles me with the possibility that my inconsistency is functional. With the perspective constituted by the prerequisite principle I can predict people with the straight edge and compass wielded by economists. With the perspective afforded by the cost principle l can predict people with the string and duct tape of psychologists. When the job resists one viewpoint, I try the other.

The zoologist Konrad Lorenz (1973, pp. 23–29) believes mood swings are functional. In an optimistic mood, I patrol for opportunities. In a pessimistic mood, I patrol for threats. Patrol requires focus, so each mood is biased. Happily, the biases cancel out over time. I am a more effective agent because of my mood swings, not despite them.

Lorenz is speculating about the advantages of diachronic inconsistency; the pendulum swings between incompatible perspectives. Changing your mind without a change of evidence is not strictly inconsistent. Each of my

temporal parts has preserved the possibility of having entirely true beliefs. But this "egoism of the moment" still reveals irrationality. At least one of my opinions lacks sufficient evidence. My father exhibited this type of irrationality when fishing on the Long Island Sound. As the tide rose, he became a more outspoken proponent of the majority view that fishing is best at high tide. As the tide ebbs, so did his allegiance to the majority. Eventually, he would defect to the minority view that fishing is actually best at low tide.

A rational agent must be prudent, reconciling the views of his temporal parts. My inconsistency about rationality is more egregious than a failure of unification. Remember that "Rationality is a scarce resource" relies on "Rationality is a prerequisite for understanding action in terms of belief and desire" and vice versa. At no moment am I consistent. Eclecticism is consistent only when choosing between consistent elements. When I wield the principle that rationality is a scarce resource, I thereby wield the contrary principle that rationality is a prerequisite. When I wield the prerequisite view, I presuppose, dialectically, the cost view.

Colin Radford (1975) argues that there are advantages to synchronic inconsistency. I can only enjoy the thrill of a horror movie if I simultaneously believe the monster is dangerous (because it is ferocious) and believe that I am safe (because there is no monster). In addition to the intrinsic pleasure of fiction, the inconsistency might make me more willing to contemplate scary scenarios. In addition to developing contingency plans, I could learn to soften stressful situations by treating them playfully, as if they were games.

I would take solace in an account of how my inconsistency is functional. But my principal hope is that you refute my self-ascription of inconsistency.

Proposal 4: Eliminate the Ingredients of Rationality!

Eliminativists about beliefs deny there are any beliefs (or desires or representations in general). They draw an analogy with souls. According to ancient philosophers, souls were essential for any psychological explanation. This led to a dilemma about how many souls to attribute. On the one hand, souls needed to be minimized to reflect the special status of human beings. On the other hand, souls needed to be maximized to account for the apparent agency of creatures that resemble human beings. Once souls are granted to women, there is pressure to extend the franchise to girls, then kittens, and so on down a slippery slope toward panpsychism. Denying that there are souls

prevents the slide at the first step. Eliminativists about souls can agree that if there are souls, then they should be maximized; and if there are souls, then they should be minimized. Since there are no souls, both the maximizing imperative and the minimizing imperative are vacuously satisfied.[3] In a similar spirit, the eliminativist agrees that if there are believers, then their rationality must be maximized and their rationality must be minimized. Since there are no believers, both imperatives are vacuously satisfied. Mission impossible becomes mission accomplished!

Eliminativists disagree about whether the eliminated item was ever possible. Paul Churchland (1981) characterizes the non-existence of belief as an empirical discovery. Neuroscientists observe and experiment with the brain and report that there is nothing corresponding to beliefs. Beliefs are possible but not actual.

A more radical eliminativist will deny that belief is possible. Just as he does not need to wait for a measurement to tell him there is no smallest fraction, he does not need to wait for an observation or experiment to show that there is no belief.

The most radical eliminativist will characterize "belief" as meaningless rather than empty (Churchland, 1986, p. 182). "There is no smallest fraction" is a meaningful theorem of arithmetic. "There is no number larger than 1/0" is not a meaningful theorem because "1/0" is ill-defined. If *belief* were well defined, then it would follow that its rationality should be both maximized and minimized. Analogy: If 1/0 were well defined, then it would equal 0 and equal infinity. So, through a glass darkly, my drive to maximize and minimize had a perverse logic. In fact, 1/0 is not a quantity at all, not even 0. So there was nothing to minimize and nothing to maximize. Since I was just manipulating symbols in a meaningless way, I was not inconsistent.

Cold comfort perhaps. But remember all I requested was a demonstration that I was not inconsistent about rationality. The most radical eliminativist has a solution that satisfies my conditions.

Eliminativism is unbelievable. Eliminativists concede this. But they deny this implication makes eliminativism self-defeating. The view that there were no souls used to be rejected as self-defeating because souls are needed for thought. The anti-eliminativist begs the question against the eliminativist.

[3] This exit became easier to see after 20th-century logicians rejected Aristotle's doctrine that "All S is P" entails "Some S is P." They satisfy the imperative "Thou shalt not suffer a witch to live" (Exodus 22:18) by virtue of the absence of witches.

But wait! If the anti-eliminativist begs the question against eliminativists, then he must have egocentrically relied upon beliefs to which he was not entitled. But then there are beliefs! The anti-eliminativist cannot consistently accuse anyone of begging the question.

Similar self-defeat awaits those who want to eliminate rationality on the basis of a cost–benefit analysis. This reduction of theorizing to trade-offs presupposes the very rationality it seeks to retire. Hard-headed philosophers picture themselves as calculating prices for theories. But the marketplace of ideas is not a theory-neutral vantage point. A philosopher has already bought into economics when he advertises himself as a detached appraiser.

Whether or not these issues of self-defeat refute eliminativism, they bear witness to the indispensability of belief. Might there be a more credible therapy for me?

Proposal 5: Treat Rationality as a Tool!

The instrumentalist is an anti-metaphysical cousin of the eliminativist. According to the instrumentalist, the aim of a belief attribution is to predict and control phenomena (Dennett, 1987). The aim is not to describe reality. Asking whether there really are beliefs misses the purpose of belief attribution. The inquirer is like a geography student who gets hung up on whether there really are lines of longitude and latitude. The correct attitude is to take the lines for granted and reckon what follows (such as great circles that minimize travel distances).

Whereas an eliminativist about Fs will admonish us for asserting or presupposing there are Fs, an instrumentalist about Fs is tolerant. He may even encourage us to continue talking about Fs with lowered inhibitions. Instrumentalists about beliefs defend beliefs as useful fictions, on par with constellations, centers of gravity, and the Coriolis force.

Instrumentalists do not worry about beliefs that cannot be specified in detail, such as the beliefs of babies, dogs, frogs, and paramecia. Since belief need only be a useful make-believe, there is no incremental danger of falling into falsehood by adding detail to the belief. Doubting that Abraham Lincoln's dog Fido has beliefs on the grounds that those beliefs would be too sketchy is like doubting that the fictional character Lassie exists on the grounds she is incomplete (since Eric Knight, the author of *Lassie Come-Home*, never specifies Lassie's weight and blood type).

A realist about belief does worry about incompleteness. For instance, Jerry Fodor (1986) denies paramecia have beliefs. But Fodor thinks frogs do have beliefs.

Stricter realists, such as Donald Davidson (1995), take the lack of specificity to show that only linguistic animals have beliefs. You believe a proposition *p* only if you have a disposition to assert *p* under conditions conducive to candor and fluency. When I report your beliefs, I say what you would say in cooperative conversation. Conversation requires give and take, with you anticipating my need to coordinate background beliefs. To attribute beliefs to you, I must picture you as able to attribute beliefs to me. To be interpreted, you must be an interpreter. So only a conversationalist can have beliefs. An attribution of beliefs and desires to speechless brutes is, at best, metaphorical.

Daniel Dennett (1987, p. 112) is leery of this literal/metaphor contrast. He adopts the intentional stance wherever it leads to successful prediction and control—even to thermostats. Thus, Dennett's principle of charity is not as substantial a commitment as Davidson's. Davidson thinks people differ qualitatively from non-linguistic animals. Dennett thinks people only differ quantitatively from frogs—belief-wise, not rationality-wise. Once you adopt the intentional stance to a frog, you have to interpret the frog as rational. People are no more rational than frogs. We adopt the intentional stance more frequently toward people because the richness of their beliefs (thanks to language) makes this strategy more predictive than for frogs.

Despite being notorious for instrumentalism about belief, Dennett (2006; Dennett & LaScola, 2013) is an outspoken opponent of religious belief. This is jarring. If belief is make-believe, what could be so bad about faith? If rationality is a prerequisite for attributing beliefs, how can the creationists be dogmatists who reason fallaciously about evolutionary theory?

By dethroning truth, instrumentalists create a power vacuum for other desiderata of belief. Beliefs become more criticizable for their origin and for their consequences. In "The Constellations Are Sexist," Leila McNeil (2016) complains that the astronomer's ways of connecting the dots in the sky echo ancient Greek and Roman mythology. Those myths reflect and perpetuate patriarchal stereotypes. Most of the figures are male, and most of those men are misbehaving. By projecting this archaic sexism into the heavens, astronomy gives a subliminal boost to poor terrestrial practices.

The instrumentalist is already doing cost–benefit analysis for belief. McNeil just continues the calculation for make-believe. She is in the same

feminist tradition that successfully lobbied for hurricanes to receive an even balance of male names and female names.

Many economists endorse instrumentalism. Milton Friedman does not worry whether each person is a rationally self-interested agent. He stresses that false axioms can systematically yield true conclusions: "theory is to be judged by its predictive power for the class of phenomena which it is intended to explain" (1953, Chapter 1, p. 8). Even false predictions are tolerable—as long as they occur outside the intended range of the theory. An instrumentalist can also tolerate inconsistency. We use calculating devices that systematically err on questions that we are unlikely to ask. The solar system model of the atom was recognized to be inconsistent with Newton's physics. But the wielders of the model used it in a way that yielded useful predictions. They did not press the model to its logical conclusion. Perhaps "Rationality is a cost" can be confirmed by measurements of increasing error under stress in the way "pi = 355/113" can be confirmed by measurements of round objects (that tiptoe around the impossible consequences of equating any rational number with pi).

Thus, the instrumentalist might agree that I am inconsistent. But he thinks that I am an alarmist. My two conceptions of rationality, even if incoherent, may be predictively useful enough to warrant continued service.

Proposal 6: Align Maximizing/Minimizing with Idealizing/Deidealizing!

I have been rescued from other apparent directional inconsistencies. When my windshield got misty, I sheepishly turned on both the defroster and the air conditioner. I did not understand why I should simultaneously maximize heat and minimize heat. I was just being faithful to a counterintuitive commandment of my car manual. Eventually, someone explained away the contradiction: Granted, no progress can be made by both heating and cooling the same body of air. But the heating and cooling apply to different bodies of air. The air conditioner removes heat from the old air (because the cooled humid air condenses and the water is expelled from the cabin). The defroster heats this dried air (thereby increasing its capacity to hold water, promoting evaporation from the windshield).

The heating/cooling inconsistency was only apparent because the transparency of air makes the newly dried air indistinguishable from old humid

air. Perhaps my maximizing/minimizing inconsistency about rationality is an analogous illusion. Instead of undoing what I did, I may be subtly cycling through to an end state.

Cycles of idealization and deidealization are portrayed this way (McMullin, 1985). Galileo's study of motion began serendipitously. A church official carrying a banner inadvertently tapped a hanging lamp. The swinging seemed intriguingly regular. Galileo measured with his pulse. As a follower of Plato, he then escaped empirical uncertainties by imagining a perfect pendulum swinging without friction in a vacuum. This thought experiment guided his laboratory construction of a U-shaped inclined plane. These experiments supported a law of equal height: The ball rolling down the ramp rolls back up the other side to the same height. Well, almost. It would were it not for the ball's friction with the plane and air resistance. This led to thought experiments subtracting this interference. Galileo eventually considered a ramp leading down to a plane of infinite length. This prevented the ball from recovering its original height. The negative lesson is that the ball continues in a straight line forever, contrary to Aristotle. The positive lesson is that an undisturbed object will continue in motion in a straight line. This anticipates Newton's first law of motion (which adds a conjunct saying that an object at rest will remain at rest).

By alternating between experiment and thought experiment, Galileo could get better and better approximations of falling objects. This cycle of idealization and deidealization is consistent because the subject matter shifts.

When trying to model information cascades, auctions, and coordination problems, economists abstract away from complexities presented by human limitations. They deal with simplified agents who are perfectly logical, have perfect memories, and are certain of the constraints that define the problem. When the model yields interesting implications, the theorists lighten the idealizations. These weakened agents are less confident about their information. They forget—or at least are uncertain about whether they will forget. Like human beings, they may lose track of time. This orderly retreat to the human condition makes the model more realistic. Experiments can be conducted in the hope of confirming that ordinary people match the model.

This interplay between idealization and deidealization suggests a strategy for resolving the dilemma between maximizing rationality and minimizing it. Associate maximizing with idealizing and minimizing with deidealizing.

This strategy could be read into L. Jonathan Cohen's (1981) import of the competence/performance distinction from linguistics. To ascertain underlying grammar, one must purge the data of mistakes due to limitations of memory, computation, and attention. A parallel purge is needed to reveal the underlying rationality of agents.

Experimenters succeed in demonstrating that people commit fallacies. But these violations of logical norms can never be evidence that human beings are logically incompetent, contends Cohen, for logical norms are based on the intuitions of normal human beings. Specifically, the logician's data are ordinary people's judgments about valid inference (just as the linguist's data are ordinary speakers' judgments of grammaticality). These data are vetted for mistakes due to inattention and memory limitations. These edited intuitions (cleansed of performance errors) become the data for reflective equilibrium. Theorists determine the strongest fit between intuitions and principles. They thereby inevitably attribute underlying competence to human beings.

The fix is in! No matter what empirical psychologists do, they get co-opted into furnishing the details of a foregone conclusion. They can only help identify the mechanism that constitutes our underlying competence.

If maximizing rationality is just minimizing irrationality (Sorensen, 1991), then delimiting the mechanism underlying competence will coincide with maximizing of rationality. Elderly people appear irrational when they fail to heed warnings. But the appearance is overturned by news that they are deaf. Perceptual mistakes do not indict rationality. By shifting the blame to perception, we restore the agent's rationality.

If one follows Thomas Hobbes in regarding memory as "decaying sense," then one can also shift blame to memory. We do not indict the rationality of the elderly for losing information they recorded on decayed paper. Blame the aging paper, not the aging reader who can no longer see the writing.

The blame shifting goes less smoothly if we view memory as heavily inferential. Reasoning is open to rational appraisal. If my false conclusion is caused by a failure to marshal evidence already under my command, then I am negligent.

Even laypeople are ambivalent about forgetting. Once you have learned an empirical fact, it exposes you to the sort of criticism associated with a priori errors such as "Socrates's aunt never had siblings." The learned fact is a given, already in your possession, in no need of further investigation. After all, when you remember a fact that slipped your mind, you have not learned

anything. When a name is on the tip of your tongue, your inability to complete the answer leaves you on the border of self-reproach.

Disturbing research on eyewitness testimony shows that we tend to be over-opinionated about how memory works (Lynn et al., 2015). So empirical work is relevant—work that will be guided by an emerging interpretative strategy. Treat the drive to maximize rationality as an artifact of the need to minimize irrationalities when identifying competence. Treat the drive to minimize rationality as a drive to find the least stock of mental resources needed to match the effect that would have been achieved with maximal rationality. Rational reconstruction can proceed in ignorance of how to realize the minimization. By specifying the goal, rational reconstruction guides the search for the minimal means.

Consider a computer that is designed to output prime numbers. Since the computer is finite, it will never finish. But this does not stop us from identifying its algorithm—despite the algorithm having infinitely many consequences.

The computer is imperfect because it is affected by electrical surges and wear and tear. Those are performance errors. These malfunctions provide clues as to the underlying program. The program is abstract. When trying to understand the program, we imagine it running a better machine, one free of memory limitations, noise, and interference.

The laws governing the computer differ from the laws of physics. According to Marvin Minsky, the very notion of a machine is value-laden:

> [Newton's mechanics] is supposed to be a generalization about some aspect of the behavior of objects in the physical world. *If the predictions that come from the theory are not confirmed, then (assuming that the experiment is impeccable) the theory is to be criticized and modified*, as was Newton's theory when the evidence for relativistic and quantum phenomena became conclusive. After all, there is only one universe and it isn't the business of the physicist to censure it, much as he might like to.
>
> For machines, the situation is inverted! The abstract idea of a machine, e.g., an adding machine, is a *specification* for how a physical object *ought* to work. If the machine that I build wears out, I censure it and perhaps fix it. (Minsky, 1967, p. 5, emphasis in original)

A description such as "The mouse is trapped" has a direction of fit that runs from words to the world. With an imperative such as "Trap the mouse!" the

direction of fit is from the world to words. Mousetraps are normative in that their direction of fit runs parallel with the mouse-catching imperative. The designer of the mousetrap gives the machine a sense of what ought to be.

You were not literally designed. But the engineering analogy still has purchase thanks to the "natural design" of natural selection. Organs are understood functionally. The heart's role is to pump blood, just as the role of a clock is to tell time.

If there is an organ of rationality (akin to the language organ posited by linguists), then rationality has functions, say accuracy in belief formation and effectiveness in satisfying desire. There is pressure to maximize these goals with the least means. Just as we cannot afford arbitrarily large hearts, we cannot afford an arbitrarily large rationality organ. We must get the most from the least (Cherniak, 1986).

If an attractive trade-off proves elusive, we may postulate another function. Hugo Mercier and Dan Sperber (2017) reject the focus on solitary reasoners. They contend that the function of reason is to persuade others through argument. It is very difficult to reason from other perspectives. So it is more efficient to reason from our perspective, leaving it to others to correct our egocentrism. Those who actually occupy other perspectives easily overcome "My-side" bias, which is so debilitating in solitary reasoning. Our mental architecture is inherited from ancestors who were hunter–gatherers, who extend their teamwork to cognitive problems. So the good news is that we are rational, as long as we are in the natural circumstances conducive to debate.

As with the previous proposals, "Maximizing is to minimizing as idealizing is to deidealizing" is too sketchy to constitute a solution. Collectively, these proposals clarify the question by roughing out a range of potential answers.

Prognosis

I am optimistic about your willingness to cure me of my inconsistency about rationality. Most people take sides on the question of whether people are irrational. Willingness to debate suggests consensus that one side is at least consistent.

I have been a cooperative patient. I have specified my feelings, findings, and failures. Indeed, I have reviewed potential cures: fanaticism, co-option, eliminativism, instrumentalism, idealization, and deidealization.

I have also tried to be a representative patient. My symptoms are common to other people. Once you cure me, you cure them. Indeed, by curing me you might cure yourself.

References

Akerlof, G. (1989). The economics of illusions. *Economics and Politics*, *1*, 1–15.

Becker, G. (1976). *The economic approach to human behavior*. University of Chicago Press.

Caplan, B. (2007). *The myth of the rational voter*. Princeton University Press.

Cherniak, C. (1986). *Minimal rationality*. MIT Press.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, *78*, 67–90.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind/brain*. MIT Press.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*(3), 317–370.

Davidson, D. (1995). Do animals have beliefs? In H. L. Roitblat & J.-A. Meyer (Eds.), *Comparative approaches to cognitive science* (pp. 111–118). MIT Press.

Dennett, D. (1987). *The intentional stance*. MIT Press.

Dennett, D. (2006). *Breaking the spell: Religion as a natural phenomenon*. Penguin.

Dennett, D., & LaScola, L. (2013). *Caught in the pulpit: Leaving belief behind*. Pitchstone Publishing.

Downs, A. (1957). *The economic theory of democracy*. Harper.

Fodor, J. A. (1986). Why paramecia don't have mental representations. *Midwest Studies in Philosophy*, *10*(1), 3–24.

Friedman, M. (1953). *Essays in positive economics*. University of Chicago Press.

Lorenz, K. (1973). *Behind the mirror*. Methuen.

Lynn, S. J., Evans, J., Laurence, J. R., & Lilienfeld, S. O. (2015). What do people believe about memory? Implications for the science and pseudoscience of clinical practice. *Canadian Journal Psychiatry*, *60*(12), 541–547.

McMullin, E. (1985). Galilean idealization. *Studies in the History and Philosophy of Science*, *16*(3), 247–273.

McNeil, L. (2016, August 16). The constellations are sexist. *The Atlantic*. https://www.theatlantic.com/science/archive/2016/08/sexism-in-the-stars/496037/#article-comments

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

Minsky, M. (1967). *Computation: Finite and infinite machines*. Prentice-Hall.

Newark Mayor Cory Booker taken to hospital after rescuing woman from house fire. (2012, April 13). *New Jersey Real-Time News*. https://www.nj.com/news/2012/04/newark_mayor_cory_booker_taken.html

Radford, C. (1975). How can we be moved by the fate of Anna Karenina? *Proceedings of the Aristotelian Society*, *49*(Suppl.), 67–80.

Samuelson, P. (1969). Presidential address the way of an economist. In P. A. Samuelson (Ed.), *International economic relations* (International Economic Association Series, pp. 1–11). Palgrave Macmillan.

Skousen, M. (1997). The perseverance of Paul Samuelson's economics. *Journal of Economic Perspectives*, *11*(2), 137–152.
Sorensen, R. (1991). Rationality as an absolute concept. *Philosophy*, *66*(258), 473–486.
Sorensen, R. (2004). Charity implies meta-charity. *Philosophy and Phenomenological Research*, *68*(2), 290–315.
Tomassello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.

2

I'm Right, You're Biased

How We Understand Ourselves and Others

Nathan N. Cheek and Emily Pronin

The desire for knowledge and understanding is a basic human motivation. Even the most rudimentary decisions that we make, such as whether to move forward to approach or backward to avoid, are rooted in an understanding, however flawed, of our ongoing state and impinging needs. When it comes to more complex decisions, such as which political candidate to vote for or what career path to pursue, we seek to make those decisions based on our understanding of our own preferences, needs, and values. Importantly, it is not only the self that we are motivated to know and understand. Knowing and understanding those around us is also a high-ranking motive. From the most basic decisions about whether to "fight" versus "flee" a potentially dangerous other to more complex decisions about whom to vote for, whom to go on a second date with, or whose advice to take seriously, we aim to make these judgments based on our knowledge and understanding not only of ourselves but of the person we are considering. But how do people come to know and understand themselves and others?

A large literature in social and cognitive psychology has revealed that people pursue knowledge about themselves and others in asymmetrical ways. When people assess themselves, they tend to *introspect* by considering internal sources of information such as goals, motives, and thoughts. In contrast, when seeking information about others, people often *extrospect* by looking instead to external information sources such as behavior. This asymmetry in strategies of information acquisition has perceptual routes: People have immediate access to their own internal states but at best limited access to the internal states of others, whereas they have direct access to others' behavior but less direct access to their own behavior (e.g., because of their visual perspective). These divergent strategies, in turn, yield divergent views of the self and others. Understanding how people attempt to learn about the

Nathan N. Cheek and Emily Pronin, *I'm Right, You're Biased* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press.
DOI: 10.1093/oso/9780197636916.003.0003

self and others can therefore shed light on why the conclusions people ultimately draw may be flawed.

In this chapter, we begin by providing a theoretical overview of how people seek information about themselves and others. Next, we highlight previous research across a variety of experimental paradigms that has documented how learning about the self through *introspection* (looking inward to thoughts, feelings, etc.) while learning about others through *extrospection* (looking outward to observable behaviors) can shape divergent ways of seeing the self and others. We then explore the mechanisms underlying asymmetrical social information–seeking strategies in the context of research on the *bias blind spot*, whereby people believe they are less biased than others. Next, turning to research on pluralistic ignorance (i.e., the widespread false belief that the group's views or feelings differ from one's own), we explore instances in which relying on extrospection can also lead people astray. Finally, we consider the implications of these epistemic approaches in the "post-truth" era.

Seeking Self-Knowledge versus Social Knowledge: Introspection versus Extrospection

Several decades of research in social psychology have documented widespread asymmetries in how people see themselves and how they see others. In their classic theoretical account, Jones and Nisbett (1972) argued that there are important differences in the evidence available to actors and observers when they seek to explain actors' behavior. Actors have direct access to their inner states, such as goals, emotional reactions, and intentions, whereas observers "have no direct knowledge of the experiential accompaniments of the act for the actor" (p. 84). In contrast, observers have direct perceptual access to actors' behaviors; indeed, Jones and Nisbett argued that "for the observer, the focal, commanding stimulus is the actor's behavior" (p. 85), whereas for actors, due to their outward visual perspective, their own behavior is less salient and accessible. The authors drew on this analysis of contrasting sources of information to explain why actors and observers make different patterns of attributions to explain actors' behavior, positing that actors' outward perceptual focus and inattention to their behavior lead them to focus on situational factors when making attributions, whereas observers'

focus on actors' behavior leads them to focus on actors' dispositions when making attributions.

Jones and Nisbett's (1972) analysis coincided with, but at least superficially seemed to contrast with, Bem's (1972) proposal that people can seek to understand themselves by using the same strategies of behavioral observation that they use to understand others. Notably, though, Bem insisted that this similarity emerges when actors' "internal cues" to explain behavior are unavailable. Nisbett and Wilson (1977) similarly proposed that people sometimes fail to find explanations for their behavior when considering inner states such as thoughts and motives and theorized that when they find introspection ineffective, people unknowingly turn to the explanatory strategies they use for others' behavior, such as lay theories.

These early and important endeavors spawned a vast literature on self-perception and social perception, much of which has considered the strategies people employ to learn about themselves and others and the consequences that arise from differential reliance on these strategies. More recent research in the vein of classic work on actor–observer differences has integrated previous theoretical approaches by demonstrating that these self–other divergences emerge out of a general tendency to mistakenly believe that introspection provides the best route to self-understanding, a belief known as the *introspection illusion* (Pronin, 2009; Pronin et al., 2004).

The introspection illusion has four components (see Table 2.1) that together describe how people differentially attend to and value diverging sources of information when they assess themselves and others (Pronin, 2009). The first component is *introspective weighting*, whereby people ascribe special status to introspection as a source of knowledge about the self. People's perceptual experience of direct access to their thoughts and feelings

Table 2.1 Components of the Introspection Illusion

Component	Description
Introspective weighting	Heavy weighting of introspection during self-perception
Self–other asymmetry	Reliance on extrospection rather than introspection during social perception
Behavioral disregard	Disregard of behavioral information during self-perception
Differential valuation	Valuing introspection as a means of self-perception but extrospection as a means of social perception

underlies the confidence they place in introspection, but this confidence is illusory because people often have little access to the causes of their behavior. Although the results of thought and judgment processes are accessible, the processes producing these results rarely are, such that introspection actually offers little explanatory insight into many thoughts, feelings, and behaviors (e.g., Bargh & Chartrand, 1999; Kahneman, 2011; Nisbett & Wilson, 1977; Pronin et al., 2002).

Second, although people view introspection as an effective path to self-knowledge, they weight internal states much less when judging others, leading to a *self–other asymmetry* in information acquisition strategies (see Figure 2.1). Instead, they rely on extrospection, drawing on others' behavior as an evidentiary source. Extrospection yields direct information from others given the perceptual access people enjoy to others' external behavior. Yet, people's own behavior is less accessible to themselves and underweighted in self-assessment, resulting in *behavioral disregard*: disregarding the self's behavior but not the behavior of others. Finally, asymmetrical reliance on

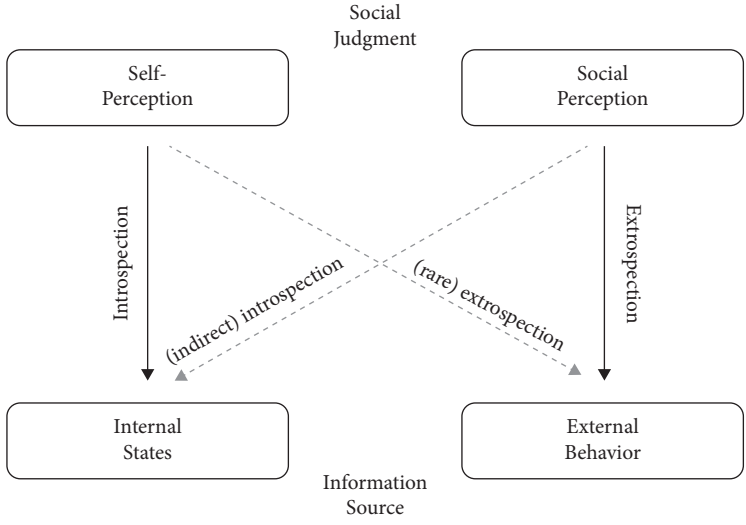

Figure 2.1 Different Information Sources for Self-Perception and Social Perception

Note: People tend to rely on introspection, looking at internal states such as intentions and goals, to learn about the self but rely on extrospection, looking at external behavior, to learn about others. Internal states are more directly available for self-perception (represented by physical proximity and a dark, solid line), whereas external behavior is less accessible for self-perception (represented by physical distance and a light, dashed line). Conversely, behavior is more directly available during social perception than others' internal states.

introspection versus extrospection arises not only from varying levels of perceptual access but also from varying levels of value attributed to these strategies. That is, the fourth and final component of the introspection illusion, *differential valuation*, describes the fact that people consider introspection to be a better source of information about the self than about others, whereas they believe that the opposite is true of extrospection (e.g., Pronin & Kugler, 2007).

The valuation of introspection as a route to self-knowledge and of extrospection as a route to other-knowledge emerges in part because of people's sense that they see the world in an objective manner. Drawing on Ichheiser (1949), Ross and Ward (1995, 1996) argued that people tend to be *naïve realists*, such that they assume that their perception of reality is veridical, unmediated by bias, ignorance, or other impeding factors. This epistemic stance leads people to trust the conclusions they reach through introspection or extrospection when assessing the self or others, respectively, because these conclusions appear untarnished by interfering factors. Thus, people mistakenly believe that they are following optimal strategies of evidence acquisition. To illustrate the breadth of this tendency, we turn now to an array of examples of self–other perceptual asymmetries that arise from the introspection illusion.

Divergent Views of the Self and Others: Some Problems with Learning from Introspection

In this section, we consider how differential reliance on introspection and extrospection leads to self–other perceptual asymmetries in judgments as varied as moral virtue, planning, free will, communication, and social influence.

Self-Righteousness

People tend to have inflated views of themselves relative to others, a tendency that has been documented across a wide variety of domains (Alicke et al., 1995; Dunning et al., 1989; Heine & Lehman, 1997). Reliance on different sources of information for judgments of the self and others greatly contributes to these overly positive self-views. For example, consider people's

tendency to believe that they are generally more moral than those around them and, in particular, that they are less likely to act immorally than even relatively similar peers (Epley & Dunning, 2000; Klein & Epley, 2017). This self-righteousness emerges because people rely on introspection when predicting their own behavior, whereas they rely on external information such as behavior and base rates to predict the behavior of others. When people consider their personal motives and thoughts, they generally find intentions to be "good"—even if their behavior does not measure up to those intentions. By contrast, people's focus on others' actions when judging others' moral behavior, without giving a lot of weight to their intentions, can instead highlight instances of less than moral behavior. In one study, college students were asked to predict their own and their classmates' future purchases of daffodils as part of a campus fraternity and sorority charity drive (Epley & Dunning, 2000). When asked whether they would buy a daffodil, 83% of participants predicted they would buy at least one flower, whereas they predicted that only 56% of other students would buy at least one flower. In reality, however, participants greatly overestimated the likelihood of buying a daffodil: Only 43% ended up making a purchase. Apparently, participants' introspection yielded not only moral intentions but also unrealistically optimistic predictions as a consequence. In a follow-up study, Epley and Dunning further found that exposing participants to information about previous donations of several peers improved the accuracy of participants' predictions of others' behavior but had no effect on predictions of their own behavior. Participants drew on behavioral information to make predictions about others but ignored it when making predictions about themselves. Further research by Helzer and Dunning (2012) has shown that when making self-predictions individuals focus on the behavior they "aspire" to rather than on the behavior they have previously engaged in.

Planning Fallacy

Using introspection to make predictions about the self also leads to the well-known *planning fallacy*, whereby people routinely underestimate—sometimes dramatically, as in the case of completing this chapter—how long it will take to complete a task (e.g., Buehler et al., 1994; Kahneman & Tversky, 1979). When estimating how long a project will take them, people focus on their positive intentions and motives while neglecting past behavioral information

(e.g., how long similar projects took them in the past) or relevant base rate information (e.g., how long similar projects typically take others) (Buehler et al., 1994; Lovallo & Kahneman, 2003). This tendency prevents people from considering possible obstacles (e.g., other work obligations) that might delay their progress and results in overly optimistic estimates. People tend not to fall victim to the planning fallacy when predicting others' performance, however, because they are more likely to take an outside view and consider previous behavior rather than industrious intentions. When participants in one study predicted the completion times of others, the estimates of those with access to information about others' previous completion times did not differ from the estimates of those with access to that information plus information about the thoughts and intentions of others (obtained through a thought-listing paradigm) (Buehler et al., 1994), indicating that they solely relied on behavioral information. Introspection (mistakenly) appears to yield reliable information for predictions of one's own completion time but is neglected in favor of extrospection when it comes to predicting others' productivity.

The Power of the Situation and Free Will

An interesting result of relying on external information when predicting the behavior of others is that people's predictions about others can be well calibrated when they correctly infer external influences on behavior. For example, Balcetis and Dunning (2013) showed that people can be relatively accurate social psychologists—in two studies, participants correctly predicted the effects of group size and mood on others' prosocial behavior. When making predictions about their own behavior, however, participants failed to take the power of the situation into account, erroneously forecasting that situational influences would not influence their prosocial behavior. By relying on their prosocial motives and positive self-views, participants discounted external factors that might influence their behavior despite their best intentions.

The findings from Balcetis and Dunning may seem surprising in light of Jones and Nisbett's (1972) famous work on the actor–observer effect. This effect is often cursorily described as a tendency for people to view their own behavior as driven by external factors ("the situation") but others' behavior as driven by internal factors ("personality"). However, a closer look at the nature of the actor–observer effect would suggest otherwise. Pronin and Kugler

(2010) suggested that individuals do not view themselves as buffeted about by the power of situational forces but rather view their actions as internally driven responses to situational forces—whereas they view others' actions as driven by unwavering internal dispositional features. This distinction leads to the hypothesis that individuals are likely to view themselves as having more free will than those around them.

A fundamental element of free will is the ability to direct one's own behavior, such that intentions and desires motivate behavior, overcoming fixed and external drivers like personality and situational forces (Watson, 1982). Because people have direct access to rich introspective content, the influence of internal states is salient when people think about their own behavior and thus seems essential to understanding when and why they act. The lack of introspective access to others' internal states, however, obscures the similar influence of motives, desires, and intentions on their behavior. Pronin and Kugler (2010) thus theorized that people may believe that intentions and desires play a larger role in driving their behavior than others' intentions and desires play in driving others' behavior, effectively ascribing more free will to themselves than others.

To test this possibility, Pronin and Kugler (2010) instructed participants to draw box models of the causes of their own behavior and that of others. Participants drew four boxes that each represented a possible cause of behavior—intentions and desires, personality, the situation, and past behavior—and then connected it with arrows to a box representing future behavior (see Figure 2.2). Participants were instructed to make the size of each box proportional to the relative influence of each cause on behavior, and the relative perceived influence of each cause was calculated by dividing the area of a given cause's box by the total area of all four illustrated boxes. Participants drew larger boxes for intentions and desires when modeling their own behavior than when modeling others' behaviors; in fact, the box for intentions and desires was drawn the largest when participants modeled their behavior, whereas the personality box was largest when participants modeled others' behavior. In other experiments, Pronin and Kugler provided further evidence that individuals viewed themselves as having more free will than others. For example, individuals asserted that there were more possible paths (both good and bad) that their lives could take than that their peers' lives could take.

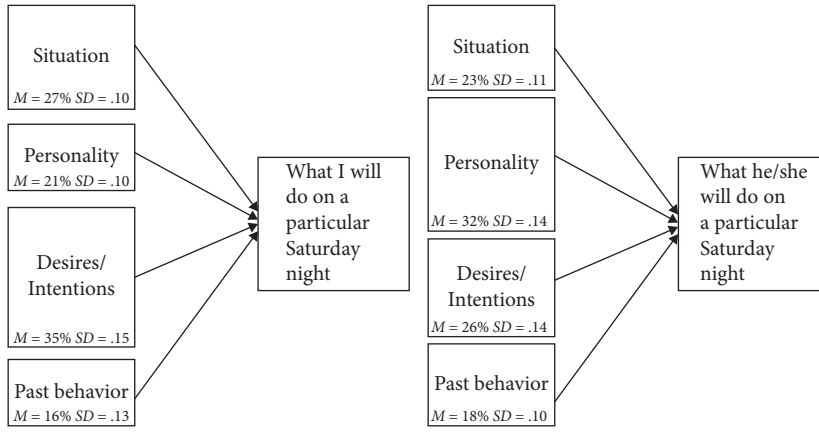

Figure 2.2 Average Images Drawn by Participants Modeling Their Own (Left) or Their Roommate's (Right) Behavior

Note: Box size is consistent with means of total area assigned to each cause of behavior.

Source: Pronin and Kugler (2010). Reprinted with permission.

Communication

In order to communicate effectively, speakers (and those communicating in other ways, through physical gestures, text messages, etc.) need to monitor the clarity of their communications and anticipate potential misunderstandings. Unfortunately, people often seem to fall short in these tasks, and they frequently overestimate how well others understand them. In one study (Keysar & Henly, 2002), for example, speakers uttering ambiguous statements overestimated how effectively they communicated an intended meaning to a listener. In a follow-up study, observers who knew the intended meaning of the statements and listened while speakers spoke were less likely to overestimate the effectiveness of communication, suggesting that it is not mere knowledge of meaning but the experience of intending to communicate it that leads to systematic overestimation of clarity in conversation. Speakers appeared to depend too heavily on their own communicative intent when gauging the comprehension of their audience, despite the importance of accurate perspective-taking in everyday conversation.

Perceptions of Social Influence

In addition to influencing predictions of future behavior or reactions, differential reliance on introspection and extrospection can shape diverging interpretations of the previous behavior of the self and others. One example concerns judgments of conformity and social influence. Pronin et al. (2007) showed that people introspect to seek evidence of possible conformity in their actions, but when conformity arises from unconscious sources (e.g., the unconscious influence of other group members' behavior), introspection yields no evidence of conformity. In contrast, extrospection often yields conformity-consistent evidence, given that conformity—by definition—results in actions looking like those of others. Hence, when people extrospect to judge whether others have conformed, they often find behavioral evidence supporting that conclusion, leading to attributions of conformity to others alongside denials of conformity in oneself.

Summary

The different strategies people use when seeking information about themselves versus others have important implications for the conclusions they reach. When people rely on introspection to learn about themselves, they may overweight their intentions and motives and underweight important external factors. Yet they rely on introspection because it feels directly accessible and accurate, more so than their behavior. Indeed, people rate their inner thoughts and emotions as more reflective of and informative about who they are than their behavior (Andersen & Ross, 1984). This introspection illusion was first identified in research on people's relative blindness to bias in themselves versus in others, and it is to this line of work we turn now.

Asymmetric Assessments and Imputations of Bias: The Bias Blind Spot

People are often blind to their own bias but quick to notice bias in others. A substantial literature has documented this bias blind spot and revealed its origins in people's asymmetric self- and social perception strategies. In initial

work on the bias blind spot, Pronin et al. (2002) found that people rate themselves as less vulnerable than others to a wide range of well-documented biases, such as the tendency to create self-serving attributions for success and failure and the tendency to selectively attend to belief-supporting evidence. This pattern emerged both when participants rated themselves relative to the average American and when they rated themselves relative to their fellow students at an elite university. Participants in Pronin et al.'s studies also denied having displayed bias immediately after displaying it in a classic "better-than-average effect" paradigm (e.g., Alicke et al., 1995; Dunning et al., 1989), and they recognized the classic self-serving bias in others' evaluations of a test they failed, despite failing to see that same bias in their own evaluations of that test.

Causes of the Bias Blind Spot

People's denial of their own bias likely has a motivational component—people are generally motivated to see themselves in a positive light, and being biased is generally seen as a shortcoming (e.g., Kunda, 1987). However, the existence of a bias blind spot is not fully explained by motivation. Indeed, in the Pronin et al. study, people did not claim to be less susceptible to other negative personal limitations, such as procrastination or fear of public speaking—if anything, they rated themselves as slightly more susceptible to these negative tendencies. The crucial difference between tendencies such as procrastination or fear of public speaking and the biases people deny is their relative cognitive availability: People are readily aware—and lament—their procrastination, whereas biases tend to operate unconsciously, leaving little trace of their occurrence. Cognitive accessibility is important because of the strategies people use when assessing bias in themselves: They introspect, searching for bias in their thought processes. This internal search yields little evidence of bias because biases operate unconsciously. Although people have access to the outputs of biased thinking (e.g., a preference for a White job applicant over a more qualified Black applicant), they lack access to the biased processes that generated those outputs (Nisbett & Wilson, 1977; Wilson & Nisbett, 1978). In contrast, when assessing the bias of others, people rely on extrospection, examining others' behavior for evidence of bias, where it can often be found.

This asymmetric reliance on introspection versus extrospection when assessing bias is not accidental: People readily report relying on introspection more when evaluating bias in the self and relying on extrospection more when evaluating bias in others (Ehrlinger et al., 2005; Pronin & Kugler, 2007). Moreover, people believe that introspection is more valuable than extrospection during self-evaluation, whereas they believe the opposite when making evaluations of others (Pronin & Kugler, 2007). This belief is based on the assumed evidentiary value of the different information-seeking strategies: People think that their thoughts would be more diagnostic of bias than their behavior, motivating the use of introspection, whereas they think that the behavior of others would be more diagnostic of bias than the thoughts of others, motivating the use of extrospection (Pronin & Kugler, 2007).

Consistent with these beliefs about evidentiary value, providing people access to the introspection of others has little effect on bias imputation. For example, Pronin and Kugler (2007) had one group of participants—*actors*—complete the same "better-than-average" task used in previous studies consisting of rating themselves relative to other students at their university on a variety of traits. These participants listed their thoughts as they completed the task and then evaluated how biased they were in their self-ratings. As expected, these participants' self-perceptions of bias were unrelated to the actual level of bias in their trait ratings. A second group of participants in this study—*observers*—assessed the amount of bias in the trait ratings of the actors and were given either only actors' trait ratings or the trait ratings along with the actors' self-reported thoughts. If observers relied on actors' thoughts when assessing their bias, then imputations of bias would differ depending on access or lack thereof to these thoughts. In reality, however, observers provided with actors' thoughts did not differ in their bias assessments from those who were not provided with actors' thoughts (which were both higher than the actors' self-assessments of bias, once again showing the bias blind spot). Observers' bias assessments did correlate with the actual level of bias in actors' behavior, however, suggesting that they were using behavioral evidence—extrospection—to evaluate bias. Thus, the bias blind spot is a problem of folk epistemology, arising from a conscious, yet flawed, belief in introspection as the best route to self-knowledge and extrospection as the best route to other-knowledge.

Breadth of the Bias Blind Spot

Blindness to one's own bias is a widespread problem. Children as young as 7 say they are less biased than other children (Elashi & Mills, 2015), and even highly intelligent people show a bias blind spot (West et al., 2012). In fact, more intelligent people may actually have a larger bias blind spot because, although they are accustomed to performing better on cognitive tasks, their cognitive sophistication does not protect them from implicit biases that arise outside of conscious awareness (West et al., 2012). Expertise also fails to shield people from the bias blind spot. For example, Neal and Brodsky (2016) interviewed board-certified forensic psychologists about potential bias (e.g., emotional connections to cases) and found that psychologists were more likely to say that other psychologists' judgments were vulnerable to bias than that their own judgments were. Moreover, all of the psychologists in the study asserted that introspection was an effective strategy for detecting bias in one's own judgments, thus revealing not only a bias blind spot but an explicit endorsement of its underlying cause.

Consequences of the Bias Blind Spot

The ubiquity of the bias blind spot has serious practical consequences for evaluations of the self and others, including insistence on the objectivity of one's own judgments, dismissal of disagreement as a result of others' bias, and the exacerbation of conflict as a result of those perceptions. First, people's use of introspection to assess their personal bias can lead them to maintain confidence in their objectivity even while recognizing their exposure to possible opportunities for bias. For example, participants in a study by Hansen et al. (2014) evaluated the quality of paintings using either an explicitly biased or an explicitly objective judgment strategy. Before rating the quality of each painting, all participants were presented with the option of learning the identity of the artist. Participants in the *objective* condition were instructed not to choose to see the artist's identity, whereas participants in the *biased* condition were instructed to choose to see the artist's identity. Participants in the latter condition explicitly acknowledged that this was a biased judgmental strategy: They rated their strategy as substantially more biased than participants in the *objective* condition. After completing the painting judgment task and using their assigned strategy to do so, participants again rated

the objectivity of their strategy, and participants in the *biased* condition again rated their strategy as less objective than participants in the *objective* condition. Importantly, however, participants in the *biased* condition did not rate their performance as more biased than that of participants in the *objective* condition. Despite acknowledging the bias inherent in their judgment strategy, they maintained that their own judgments were objective; in fact, their confidence in their own objectivity increased after knowingly using the biased strategy.

People's confidence in their objectivity after using a biased strategy stems from their use of introspection: When people examine their thoughts and motives for traces of bias, they find no evidence of bias. Moreover, because they know their strategy was biased, this lack of evidence is even more remarkable—it suggests that they have maintained objectivity in the face of biasing influences. Hence, people may feel even more confident in their objectivity after employing a biased judgment strategy than before employing it. This pattern of self-perceived objectivity as a result of introspection may explain the aforementioned confidence that forensic psychologists feel despite knowing that biases such as emotional connections to defendants could potentially influence their professional judgment. It also underlines the limitations of attempts to warn people about potentially biasing factors. For example, jurors exposed to biasing testimony and instructed to ignore it may feel they are preventing it from biasing their judgments while nonetheless being influenced. Similarly, journal reviewers may be confident that knowing the identity of a manuscript's author will not sway their conclusions, despite their vulnerability to, for example, favoring authors from prestigious universities. Reliance on introspection undermines people's ability to judge bias in themselves, leading them to maintain—and perhaps even gain—confidence in their objectivity as they interpret an absence of evidence of bias to indicate evidence of an absence of bias.

Introspection and the resulting bias blind spot can also cause people to dismiss disagreement; rather than taking disagreement seriously and reexamining their own views, they attribute it to the bias of others. In a study by Kennedy and Pronin (2008), for example, participants read a fictional article about the then-new president of Harvard that included a discussion of her views on affirmative action. Her views were presented as relatively moderate, and there was large variability in participants' own affirmative action opinions, resulting in varying levels of disagreement between participants and Harvard's president. After reading the article, participants rated their

perception of how objective or biased the president was, and these ratings were shaped by whether they agreed or disagreed with her—the more they disagreed with her stance, the more they attributed her views to bias rather than careful reasoning. Kennedy and Pronin found the same pattern of results when manipulating disagreement, providing causal evidence that people respond to disagreement by attributing conflicting viewpoints to bias.

People react to disagreement by imputing bias to others in part because introspection reveals no evidence of bias in their own views; thus, the source of disagreement seems as though it cannot be one's own bias. The view of oneself as objective precedes disagreement, however. A central tenet of naïve realism is that people believe they see the world as it is. And so it goes, when others see things differently, people are left to infer that those others are either ignorant or misinformed—or biased (Ross & Ward, 1995, 1996; see also Ichheiser, 1949). Thus, folk epistemology devalues disagreement as largely a reflection of a disagreeing other's faults, whether it be their ignorance, misinformation, or—having ruled out those alternatives—bias. And discounting disagreeing others as a potential source of information has negative implications for information acquisition and knowledge—Liberman et al. (2012), for example, found that participants failed to incorporate the judgments of disagreeing others in an incentivized judgment task and earned less money based on performance because they dismissed the informational value of opposing judgments.

Devaluing disagreement is not only problematic from a rational epistemic standpoint; it can also incite and escalate conflict, resulting in a bias-perception conflict spiral. Disagreement can lead to conflict, but not all disagreement follows such a negative path. One important catalyst that moves a disagreement between two parties toward conflict is uncooperative, competitive behavior toward the other side. Kennedy and Pronin (2008, 2012) showed that people react to disagreement by imputing bias to opposing parties. They also found, however, that imputing bias to opponents leads people to react more competitively and aggressively. To some extent, this latter reaction is understandable and, in some cases, could be rational—a biased, unreasonable opponent is unlikely to respond well to cooperative conflict-reduction strategies, such that a more competitive response seems more effective. Importantly, though, people react more competitively not just in response to perceived bias but in response to mere disagreement because they take the extra inferential step of attributing disagreement to bias. This response produces a conflict spiral because people's uncooperative

responses in the face of disagreement (and inferred bias) are then seen as aggressive and biased by the opposing party, producing a competitive response in return. Thus, imputation of bias in the face of disagreement risks inciting a conflict spiral that unnecessarily escalates a situation that could perhaps be resolved more cooperatively.

Reducing the Bias Blind Spot

Given the problems posed by the bias blind spot, a pressing question is, how can the bias blind spot be reduced? In this section, we highlight three potential strategies to help overcome the failure to see bias in oneself. One approach to reducing the bias blind spot is "exposure control" (Gilbert, 2002; Wilson & Brekke, 1994)—that is, trying to prevent bias from occurring in the first place by removing the presence of biasing influences. Double-blind peer review, for example, attempts to remove the possible influence of author identity to prevent bias from taking place. This approach has strong merits; after all, the problem of bias blindness only exists when bias itself exists; thus, eliminating the opportunity to be biased also eliminates the problem of bias blindness. In many situations, however, it may be impossible or unfeasible to fully remove biasing influences, and thus other approaches are needed as well.

A tempting additional approach to reducing the bias blind spot would be to simply educate people about implicit biases, emphasizing that they are common, are frequent, and can occur outside awareness. The ironic problem posed by the bias blind spot, however, is that it results from the failure to recognize bias in oneself; hence, educating about bias cannot alone reduce it. Indeed, people may well respond to education about biases by noting how frequently they see those around them display such tendencies, while still neglecting those same tendencies in themselves as a result of relying on introspection. Thus, the topic of education should not just be bias itself but rather the limitations of introspection as a strategy to assess bias. Pronin and Kugler (2007) tested such an approach by giving one group of participants a (fabricated) article about the limits of introspection, which reviewed actual psychological research about how people are unable to detect their own bias by looking within. A second group of participants instead read an unrelated article about pollution. When participants then rated their own susceptibility to a series of biases relative to the susceptibility of other undergraduates

at their university, those who read the pollution article displayed the classic bias blind spot, but those who learned about the flaws of introspection did not. Thus, education can reduce the bias blind spot if focused on the strategies people use to seek information about their own bias.

A third possible approach changes the focus from introspection to extrospection by encouraging people to focus on outward behavior. That is, instructing people to focus on the appearance of bias in their own behavior—rather than internal indicators of bias—can make people more likely to acknowledge the possibility of bias. When employees are asked to complete conflict of interest reports, the focus is often on disclosing anything that might appear biased. Even federal judges, who are expected to be beacons of objectivity, are required to recuse themselves from a federal case if a reasonable person might question their impartiality. These standards essentially ask people not to look inward for evidence of objectivity but rather to look outward to how others might see them (though their views of what would look like bias may derive from internal standards rather than others' standards).

In summary, the bias blind spot is a consequential asymmetry in the perception of the self and others and largely results from different information-seeking strategies. People readily notice bias in others by focusing on others' behavior and lay theories of bias, whereas they fail to notice it in themselves because they rely on introspection. Introspection, despite feeling like the most effective self-assessment strategy, is ineffective because people have little access to thought processes, preventing them from successfully detecting bias in themselves.

The Sometimes Problem with Looking at Behavior: Pluralistic Ignorance and the Deviance Assumption

Although people do not have introspective access to their thought processes, people do have introspective access to the contents of their thoughts, as Nisbett and Wilson famously pointed out in their 1977 paper. Thus, introspection is not as problematic when people seek self-knowledge that is consciously available in thought contents. On the other hand, although extrospection is sometimes helpful in the detection of bias in others—because bias can leave behavioral traces even when its tracks are covered in consciousness—extrospection can also be problematic. It can lead to

information-acquisition problems of its own when the behavior of others is incorrectly assumed to reflect their internal thoughts, feelings, knowledge, and experiences.

Perhaps the best-known case of when extrospection yields inferior information to introspection is what Floyd Allport (1924) termed *pluralistic ignorance* (see also Prentice & Miller, 1994, 1996). Pluralistic ignorance occurs when people mistakenly believe that an unpopular social norm is widely endorsed by the group. The phenomenon arises as a consequence of the fact that social information must often be indirectly acquired through the observation of others' behavior (and this is particularly true when people seek information about the group rather than an individual). This route to knowledge can be problematic, though, when people's public behavior fails to accurately represent their private opinions. In those cases, extrospection (i.e., relying on observable behavior) can lead to drastically incorrect conclusions about social norms, resulting in pluralistic ignorance.

In a classic study, Prentice and Miller (1993) surveyed Princeton University undergraduates about their level of comfort with the drinking habits of Princeton students, as well as how comfortable they thought the average Princeton student was. Most participants reported that they were much less comfortable with campus drinking habits than the average Princeton student, such that the actual average level of comfort with drinking habits was much lower than the perceived level of comfort. Thus, participants mistakenly believed they differed from their peers and perceived an unpopular norm—heavy drinking—as popular. Prentice and Miller's research demonstrates that people can fail to acquire important social knowledge. Indeed, norms are important drivers of behavior, either because people want to fit in and therefore conform to perceived norms or because people use norms as information about how they should behave. It is precisely the tendency to conform to norms, however, that creates the potential for pluralistic ignorance—people want to fit in, and thus they publicly act in accordance with what they perceive as normative. As a result, most people outwardly act in accordance with the perceived norm, despite not privately endorsing it. When people seek to learn about collective opinions through extrospection, the behavioral evidence they encounter suggests that others endorse the unpopular norm.

That participants knowingly misrepresent their private views in public but nonetheless use others' public behavior to infer those others' private views highlights a pitfall of relying on extrospection for others—people seek

information through others' behavior even while knowing that their own behavior is not dispositive of the internal motives that are guiding that behavior. Similar to the way that relying on introspection to assess one's own bias induces a bias blind spot, relying on extrospection to assess others' internal states can create a deviance assumption or "normalcy blind spot," whereby people are blind to the extent to which they are similar to others—that is, "normal."

Relying on introspection can create the false impression that one is deviant or abnormal, in a socially undesirable way, particularly because public behavior—the evidence on which extrospection draws—often conceals negatives and showcases positives. For example, Jordan et al. (2011) found that people erroneously thought they experienced more negative emotions than others, despite also recognizing that they are more likely to conceal negative emotions and display positive ones. Indeed, the extent to which peers reported concealing negative emotions predicted the degree to which individuals thought they experienced abnormally high levels of negative emotions. Moreover, participants who underestimated others' negative emotional experiences to a greater extent experienced more loneliness and decreased subjective well-being. Thus, people's blindness to their emotional normalcy created a false impression of being worse off than those around them.

In a related study, Cheek and Pronin (2018) found that 225 adults in a Mechanical Turk sample believed that even when they were knowingly putting on a happy face at a party despite not actually enjoying a party, their peers' happy faces were more likely to reflect actual enjoyment. Specifically, participants were told to imagine the following situation:

> You are at a work-related party on a Friday night, and everyone is milling around with a glass of wine in their hand. During a casual conversation with various party-goers, someone asks the group of you if they are enjoying themselves. You are exhausted and just want to go home. Nonetheless, fear of seeming like you're no fun prevents you from answering honestly. You happily nod your head yes and have another sip of wine. You notice that everyone else is smiling happily too. Why do you think everyone is smiling happily in response to the question?

Participants then rated how likely it was that the other guests were enjoying the party and how likely it was that the other guests were smiling to give off

a good impression on scales from 1 to 7, with higher numbers indicating greater likelihood. Even though they knew their own smile was merely a mask, participants thought it was much more likely that others were smiling out of genuine enjoyment (M = 5.29, SD = 1.31) than that others were pretending to enjoy the party (M = 3.96, SD = 1.67), $t(224) = 7.65$, $p < .001$, $d = .51$ (see Figure 2.3). Thus, people interpreted the behavior of others as honestly reflecting inner states, despite their introspective knowledge that belied their own smiles. Extrospection as a means to understand others' thoughts and feelings is only valuable insofar as behavior provides a reasonable proxy for inner states, and reliance on extrospection to learn about others' inner states may be so valued that people do not apply their knowledge of its limitations in even relatively transparent cases.

Pluralistic ignorance and the normalcy blind spot pose problems for the collective as well as for individuals. Unpopular norms will survive due to pluralistic ignorance, leading people to conform to them long after they are endorsed by few. Kugler and Darley (2012), for instance, found that people tend to overestimate others' agreement with and support for current drug

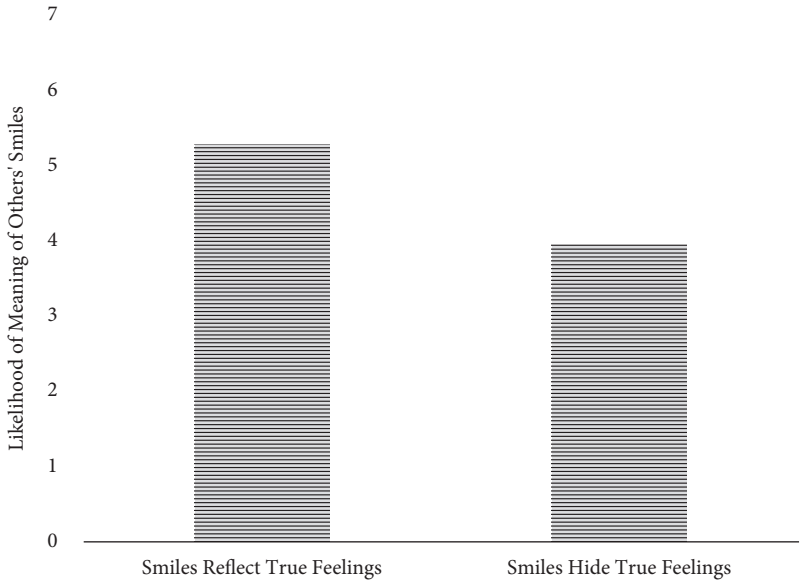

Figure 2.3 Beliefs About Informativeness of Extrospective Evidence in a Pluralistic Ignorance Situation

Note: Participants thought it was more likely that others' smiles reflected their true feelings, despite knowing that their own smiles hid their true feelings.

enforcement policies, presumably because public outcry against those policies has been limited. Public behavior that misrepresents private views can impact policies and laws in addition to individual behavior by leading those in power to misunderstand public opinion.

Reducing Pluralistic Ignorance and the Normalcy Blind Spot

The normalcy blind spot results from both the fact that people misrepresent their internal states in public and the fact that people use extrospection to assess others' internal states. Thus, one strategy to reduce pluralistic ignorance and the normalcy blind spot is to educate people about others' actual private views and experiences. Making private information known and available can decrease the deviance assumption and reduce pluralistic ignorance. For example, communicating rates of college students' seeking treatment for depression and anxiety or of their experiencing discomfort with casual sex on campus may help reduce individual students' feeling that they are alone in having anxiety or depression or that they are uniquely uncomfortable with casual sex. In the political realm, exposing citizens to the results of opinion polls can provide a novel source of information that is more accurate than extrospection.

An alternative strategy is to educate people about the limits of extrospection as a means to assess others' true inner states (and, consequently, as a means to draw conclusions about one's own deviance). This approach parallels the strategy of reducing the bias blind spot by educating people about the limits of introspection (e.g., Pronin & Kugler 2007). For example, Schroeder and Prentice (1998) found that undergraduates assigned to participate in a discussion of pluralistic ignorance reported less drinking several months later relative to undergraduates assigned to a control discussion. Thus, pluralistic ignorance and the normalcy blind spot can be addressed either by attempting to provide a better source of evidence of others' inner states or by attempting to expose the limits of relying on extrospection to seek out evidence of others' inner states.

Perceptions of Bias in the "Post-Truth" Era

Psychological and philosophical approaches to understanding how people gain knowledge—and how they should gain knowledge—about the social

world are perhaps more important than ever as we find ourselves in the "post-truth" era. Ample research has found that people view their own perceptions as reflecting objective reality while they readily impute bias to those who disagree with them. Armed with this psychological tendency, people can discount disagreements as the result of other people's bias or ignorance, rather than take disagreements seriously as an opportunity to learn new information, update beliefs, and achieve better understanding and mutual cooperation.

Indeed, journalists, academics, and cultural commentators alike have decried the recent ascendance of unfounded beliefs, "fake news," and "alternative facts" in modern politics (e.g., d'Ancona, 2017; Davies, 2016, 2018; Manjoo, 2008; McIntyre, 2018; Rabin-Havt & Media Matters for America, 2016). One broad theme in reflections on the post-truth era is an apparent disregard for veracity—a motivated denial of facts in favor of other, more appealing "alternative" ones. As psychology has long recognized, motivation can shape how people seek out information and the conclusions they draw from it (e.g., Ditto & Lopez, 1992; Dunning et al., 1989; Kunda, 1987); and as the internet facilitates access to a seemingly infinite number of potential information sources, it is easier than ever to create an ideological bubble that confirms one's worldviews to the detriment of a broader understanding of reality. But even beyond motivated distortion of the truth, the research reviewed in this chapter underlines social and cognitive limitations in the search for social information, rooted in the introspection illusion.

We cannot expect people to gain introspective access to the biases that distort their perceptions and to see the action of those processes bending their perceptions away from "objective reality" and toward an alternative one. Those distorting processes happen unconsciously, thereby affording people an undue confidence in their own objectivity and correctness. We also cannot expect people to gain an appreciation for others' grasp on objective reality. After all, those others lack that grasp as much as we do—and we therefore are correct in claiming that others are biased (even if we are incorrect about the amount of their bias). These things are unreasonable to expect. What we can strive for, however, as people who value a fair and just society characterized by harmony rather than by strife, is to entertain the possibility that we too are biased. We can strive to look at our own behaviors when we look at those of others, and when those behaviors seem like signs of bias in others, we can acknowledge that the same behaviors in ourselves may similarly signal bias. Our internal, introspective worlds can be rich, beautiful, and

revealing. But relying on them while judging others from the outside can also be a costly divergence.

References

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology*, *68*(5), 804–825.

Allport, F. H. (1924). *Social psychology*. Houghton Mifflin.

Andersen, S. M., & Ross, L. (1984). Self-knowledge and social inference: I. The impact of cognitive/affective and behavioral data. *Journal of Personality and Social Psychology*, *46*(2), 280–293.

Balcetis, E., & Dunning, D. (2013). Considering the situation: Why people are better social psychologists than self-psychologists. *Self and Identity*, *12*(1), 1–15.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462–479.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). Academic Press.

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *67*(3), 366–381.

Cheek, N. N., & Pronin, E. (2018). *Pluralistic ignorance and naïve realism* [Unpublished raw data].

d'Ancona, M. (2017). *Post truth: The new war on the truth and how to fight back*. Ebury Press.

Davies, W. (2016, August 24). The age of post-truth politics. *The New York Times*. https://nyti.ms/2bz5Jr5

Davis, E. (2018). *Post-truth: Why we have reached peak bullshit and what we can do about it*. Little, Brown.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*(4), 568–584.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, *57*(6), 1082–1090.

Elashi, F. B., & Mills, C. M. (2015). Developing the bias blind spot: Increasing skepticism towards others. *PLoS ONE*, *10*(11), Article e0141809.

Epley, N., & Dunning, D. (2000). Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social perception? *Journal of Personality and Social Psychology, 79*(6), 861–875.

Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot: People's assessments of bias in themselves and others. *Personality and Social Psychology Bulletin*, *31*(5), 680–692.

Gilbert, D. T. (2002). Inferential correction. In T. Gilovich, D. W. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 167–184). Cambridge University Press.

Hansen, K., Gerbasi, M., Todorov, A., Kruse, E., & Pronin, E. (2014). People claim objectivity after knowingly using biased strategies. *Personality and Social Psychology Bulletin*, *40*(6), 691–699.

Heine, S. J., & Lehman, D. R. (1997). The cultural construction of self-enhancement: An examination of group-serving biases. *Journal of Personality and Social Psychology*, *72*(6), 1268–1283.

Helzer, E., & Dunning, D. (2012). Why and when peer prediction is superior to self-prediction: The weight given to future aspiration versus past achievement. *Journal of Personality and Social Psychology*, *103*(1), 38–53.

Ichheiser, G. (1949). Misunderstandings in human relations: A study in false social perception. *American Journal of Sociology*, *55*(Suppl.), 1–70.

Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the cause of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79–94). General Learning Press.

Jordan, A. H., Monin, B., Dweck, C. S., Lovett, B. J., John, O. P., & Gross, J. J. (2011). Misery has more company than people think: Underestimating the prevalence of others' negative emotions. *Personality and Social Psychology Bulletin*, *37*(1), 120–135.

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

Kahneman, D., & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. *TIMS Studies in Management Science*, *12*, 313–327.

Kennedy, K. A., & Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin*, *34*(6), 833–848.

Kennedy, K. A., & Pronin, E. (2012). Bias perception and the spiral of conflict. In J. Hanson (Ed.), *Ideology, psychology, and law* (pp. 410–446). Oxford University Press.

Keysar, B., & Henly, A. S. (2002). Speakers' overestimation of their effectiveness. *Psychological Science*, *13*(3), 207–212.

Klein, N., & Epley, N. (2017). Less evil than you: Bounded self-righteousness in character inferences, emotional reactions, and behavioral extremes. *Personality and Social Psychology Bulletin*, *43*(8), 1202–1212.

Kugler, M. B., & Darley, J. M. (2012). Punitiveness towards users of illicit drugs: A disparity between actual and perceived attitudes. *Federal Sentencing Reporter*, *24*(3), 217–221.

Kunda, Z. (1987). Motivation and inference: Self-serving generation and evaluation of evidence. *Journal of Personality and Social Psychology*, *53*(4), 636–647.

Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the "wisdom of dyads." *Journal of Experimental Social Psychology*, *48*(2), 507–512.

Lovallo, D., & Kahneman, D. (2003). Delusions of success: How optimism undermines executives' decisions. *Harvard Business Review*, *81*(7), 56–63.

Manjoo, F. (2008). *True enough: Learning to live in a post-fact society*. John Wiley & Sons.

McIntyre, L. (2018). *Post-truth*. MIT Press.

Neal, T. M. S., & Brodsky, S. L. (2016). Forensic psychologists' perceptions of bias and potential correction strategies in forensic mental health evaluations. *Psychology, Public Policy, and Law*, *22*(1), 58–76.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, *64*(2), 243–256.

Prentice, D. A., & Miller, D. T. (1994). Collective errors about the collective. *Personality and Social Psychology Bulletin*, *20*(5), 541–550.

Prentice, D. A., & Miller, D. T. (1996). Pluralistic ignorance and the perpetuation of social norms by unwitting actors. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 28, pp. 161–209). Academic Press.

Pronin, E. (2009). The introspection illusion. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 41, pp. 1–67). Academic Press.

Pronin, E., Berger, J., & Molouki, S. (2007). Alone in a crowd of sheep: Asymmetric perceptions of conformity and their roots in an introspection illusion. *Journal of Personality and Social Psychology*, *92*(4), 585–595.

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the eye of the beholder: Divergent perceptions of bias in self versus others. *Psychological Review*, *111*(3), 781–799.

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *43*(4), 565–578.

Pronin, E., & Kugler, M. B. (2010). People believe they have more free will than others. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(52), 22469–22474.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381.

Rabin-Havt, A., & Media Matters for America. (2016). *Lies, incorporated: The world of post-truth politics*. Anchor.

Ross, L., & Ward, A. (1995). Psychological barriers to dispute resolution. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 27, pp. 255–304). Academic Press.

Ross, L., & Ward, A. (1996). Naïve realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed & E. Turiel (Eds.), *Values and knowledge: The Jean Piaget Symposium series* (pp. 103–135). Erlbaum.

Schroeder, C. M., & Prentice, D. A. (1998). Exposing pluralistic ignorance to reduce alcohol use among college students. *Journal of Applied Social Psychology*, *28*(23), 2150–2180.

Watson, G. (1982). *Free will*. Oxford University Press.

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*, *103*(3), 506–519.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117–142.

Wilson, T. D., & Nisbett, R. E. (1978). The accuracy of verbal reports about the effects of stimuli on evaluations of behavior. *Social Psychology*, *41*(2), 118–131.

4

Can We Be Reasonable?

Bias, Skepticism, and Public Discourse

Teresa Allen and Michael Patrick Lynch

A long-standing and influential thought is that for democracies to function well—or perhaps to function at all—they need vigorous but reasonable public discourse. The ideal is that they should be spaces of reasons—spaces where reasons for policy decisions can be exchanged and listened to. Yet there is mounting evidence suggesting that not only are human beings subject to biases and errors in reasoning but we are also particularly bad at spotting when they are affecting us. As a result, one might suspect that we should be deeply skeptical about whether public discourse can ever be reasonable.

In this chapter, we follow this suspicion to its logical conclusion, raising a novel skeptical argument based on the problem of what we'll call "bad bias." This skeptical argument, we believe, raises a serious challenge to the possibility of reasonable public discourse. Even so, reflection on the argument also points us toward new ways of confronting this challenge—a challenge that arguably goes to the heart of democracy itself.

Reasonableness as a Norm of Public Discourse

Reasonableness is often understood as a property of beliefs, statements, arguments, or theories. However, we'll be taking it here as a name for a norm or ideal of public discourse. According to this ideal, public discourse ought to involve the exchange of reasons that participants (i.e., those who wish to engage in public discourse) can recognize as reasons.[1] Discourse, as well as

[1] Here, the emphasis should be placed on the *can*, for participants might not recognize reasons offered by others as such; alternatively, they may mistakenly consider certain things to be reasons which are not in fact reasons. The point is that participants in a democracy must be capable of

Teresa Allen and Michael Patrick Lynch, *Can We Be Reasonable?* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press. © Oxford University Press 2022.
DOI: 10.1093/oso/9780197636916.003.0005

particular exchanges within it, is reasonable to the degree that participants meet this ideal.

Thus stated, the *norm of reasonableness* is a familiar element of so-called deliberative conceptions of democracy. Broadly speaking, these conceptions take democracies to be spaces of reasons—spaces where conflicts between citizens can be worked out not only at the ballot box but also via the exchange of reasons and ideas.[2] It lays down two main requirements: that discourse involves an exchange of reasons and that the reasons in question be—at least in principle—recognizable as reasons by the different participants in the discourse.

The rationale behind the first requirement is fairly obvious. The point of public discourse is generally to achieve some mixture of knowledge, consensus, or—at the very least—clarity regarding matters of policy and law. As a result, we presumably want the beliefs we defend in, or form as a result of, public discourse to be justified. If reasons were not traded—but instead, say, only threats, intuitions, emotional reactions, ad hominem attacks, or the like—then it would be hard to imagine how beliefs defended (or formed) about the effectiveness of certain policies and laws would be justified. Without appeal to reasons, our ability to obtain political knowledge, at least via public discussion, would be greatly diminished.

The second requirement, though perhaps not as obviously true, is still plausible for at least two reasons.[3] First, if the goal of discourse involves increased consensus and clarity about policy, simply announcing one's reasons to the world is not very helpful. Depending on how they are packaged, reasons might appear to be merely intuitions or emotional appeals. Alternatively, they might presuppose a worldview others don't share or knowledge and experience that others lack. In order for participants to successfully defend their beliefs in the realm of public discourse, it helps if—indeed, it is arguably necessary that—they use reasons others can recognize as moves in the same basic game. Second, democracy requires that citizens treat each other with basic forms of respect—in particular, as beings that can, at least in principle, make up their own minds. Giving others reasons that they could—should

recognizing the (actual) reasons offered by others, even if they do not in fact recognize them as such. For more discussion, see Lynch (2016, especially p. 252).

[2] For classic expressions of such views, including the different ways in which the norm of reasonableness can be understood, see Rawls (1971), Habermas (1985), Anderson (2006), and Talisse (2009).

[3] For further discussion and a defense of the following point, see Lynch (2012); for an alternative account of public reason more generally, see Gaus (2012).

they put their minds to it—recognize as reasons treats them in this way. Giving others reasons they could never appreciate presumably does not.[4]

In short, both considerations suggest that trading reasons in the realm of public discourse is much like trading currency. Currency no one recognizes as such is no currency at all. The same goes, one might think, for reasons traded in the arena of public discourse: Their value is retained only if they are recognizable as reasons.

Even so roughly described, the norm of reasonableness is remarkably difficult to meet. This in itself may not be particularly surprising—such is the nature of ideals. But if we want to understand how that ideal affects our practices and how we can even make progress toward meeting it, we need to understand the demands the norm places upon us. Thus, we begin by laying out several conditions, starting with two demands the ideal of public discourse imposes on citizens' cognitive architecture. First, there is

> *Reflection*: Participants must have the cognitive capacity to recognize reasons exchanged within public discourse as reasons.

Not only must those who offer reasons ensure that the reasons are recognizable but also, those who are presented with the reasons must have the capacity to recognize them as such.[5] In other words, reflection tells us that participants in reasonable public discourse must be able to separate the epistemic wheat from the chaff.

In addition to the first condition, we need

> *Response*: Participants must be willing and able to revise their cognitive states in response to reasons.

Recognizing reasons as such will do little good in the realm of public discourse if nobody changes their minds as a result of being presented with good reasons. In order for public discourse to be healthy, participants must be responsive to reasons offered by others.

Much of the discussion in this chapter will concern reflection and response since focusing on them brings out the difficulty of meeting the norm

[4] A further question—not to be addressed here for lack of space—concerns how to understand the testimony of experts in democratic life. For reflections on this question, see Goldman and Whitcomb (2011) and Lynch (2012).

[5] For more on reflection, see Kornblith (2012).

of reasonableness. However, they are not the only demands that the norm lays down. Others—the importance of which will become apparent toward the end of this chapter—include

> *Cognitive Empathy*: Participants must possess the imaginative capacity to see where others are coming from.[6]

Each participant in public discourse holds beliefs that are—presumably to a large extent—shaped by their overall perspective, where that includes their concepts, beliefs, experiences, and social identity. In order for recognizable reasons to be exchanged, participants must be capable of imagining occupying others' perspectives and deliberating from the principles recognized within those perspectives.[7]

However, to meet the norm of reasonableness it is not enough that participants be responsive, reflective, and cognitively empathetic; for if they are so epistemically arrogant that they never recognize the reasons offered by others as ones that could in any way improve their epistemic situations, then it makes little difference whether they are responsive, reflective, and cognitively empathetic—their beliefs would never change anyway. To see why, suppose a participant is able to recognize others' reasons as such, that they are able to put themselves in their epistemic shoes, as it were, and that they are even able and willing to revise their cognitive states in response to reasons. However, because they do not view the reasons provided by others as all that good—at least not better than the reasons they possess—and because putting themselves in the epistemic shoes of others does not change anything for them, then their cognitive states are rarely, if ever, affected as a result of considering others' reasons or experiences.[8] This motivates the need for the following condition:

[6] Importantly, possessing the imaginative capacity to see where others are coming from is not the same as (or as strong of) a requirement as being capable of knowing what it's like to be someone else. For more on the limitations of our ability to imagine certain kinds of experiences, see Paul (2014).

[7] For more on what it means to imagine occupying others' perspectives, see Darwall (2013, Chapter 6).

[8] It is an open question whether such an arrogant person can really count as being responsive. Likewise, it is an open question as to whether they are really capable of being cognitively empathetic if they are unmoved by putting themselves in the shoes of others. Be that as it may, we might assume for the sake of argument that they would revise their beliefs in response to the reasons and experiences of others—that is, if only the reasons offered were "good enough" or experiences "compelling enough" for their (skewed) standards. The point is simply that the aforementioned conditions are at least necessary, but not obviously sufficient, for meeting the norm of reasonableness.

> *Intellectual Humility*: Participants must be both willing and able to see their epistemic situations as capable of improvement.[9]

If this condition is not met, it is hard to see how participants in a democracy would ever accept others' reasons as ones they themselves should adopt or others' experiences as ones from which they might have something to learn.[10]

Finally, to meet the norm of reasonableness, the following condition is needed:

> *Epistemic Equality*: The background epistemic system must equally recognize participants' capacities for reason-giving, it must allow for equal access to epistemic resources, and it must not contribute to cognitive marginalization.

The rationale for this condition is straightforward: Public discourse will suffer if it is infected by epistemic injustice.[11] Not only will those who are epistemically marginalized suffer as their recognizable reasons will not be recognized—for example, because those who marginalize them fail to be sufficiently reflective—but also those who are not marginalized lose out on the valuable contributions that would otherwise be made by those who are marginalized. For a public discourse to flourish, the reasons of all—and not just the reasons of some—must be recognized and respected (Fricker, 2007; Medina, 2012, 2018).

Before moving on, two points are worth mentioning. First, cognitive empathy and intellectual humility—like reflection and response—are personal conditions in the sense that they concern the cognitive capabilities of ordinary citizens. Epistemic equality, on the other hand, is an institutional demand. Although individual citizens can promote it, epistemic equality cannot be realized by the efforts of individual citizens alone; rather, it must be realized by citizens as a collective. Second, meeting the norm of reasonableness is

[9] This is compatible with holding that they ought to stick to their guns when their reasons are objectively good (and this is something they can recognize).

[10] For more discussion on the concept of intellectual humility, see Whitcomb et al. (2015), Tanesini (2016), and Lynch (2018b).

[11] Following Miranda Fricker (2007), we understand *epistemic injustice* to be "a kind of injustice in which someone is *wronged specifically in her capacity as a knower*" (p. 20, original emphasis). Importantly, the core kinds of harms picked out by the concept of epistemic injustice are harms that epistemic agents suffer in virtue of their (actual or merely perceived) social identity. Thus, for example, women suffer from epistemic injustice when—because they are women—their intellectual contributions are viewed as less valuable than their male counterparts' contributions.

a matter of degree, and discourse can be more or less reasonable by meeting more or fewer of these conditions. Moreover, meeting, or promoting the realization of, each of these conditions themselves can be a matter of degree. For example, participants may be better at being responsive and epistemically humble than they are at promoting epistemic equality—they might see their epistemic situations as capable of improvement and be capable of revising their cognitive states in response to reasons but only if the reasons are offered by certain participants. Or perhaps they are better at being cognitively empathetic than they are at being responsive. In any case, here is the takeaway point: Since participants may be better or worse at meeting, or promoting the realization of, these conditions, public discourse will consequently be more or less reasonable depending on how well participants meet these conditions, both on balance and individually.

Against Reflection and Response

Assuming we are right that public discourse is reasonable to the degree it meets these conditions, it is sensible to ask whether, and to what degree, these conditions can actually be met. As noted, we'll concentrate on the conditions of response and reflection.

Given the present state of much political discourse, one could not be blamed for thinking that the prospects of meeting the norm of reasonableness are rather bleak. But one can do more than simply speculate on the matter, for there is a vast and ever-growing body of research which both exposes our flaws as reasoners and sheds light on just how ineffective we are when it comes to recognizing that our reasoning is flawed. Taken together, this evidence suggests that meeting reflection and response is actually difficult, which in turn suggests that meeting the norm of reasonableness is difficult as well. We begin by listing some of the relevant research and then turn to how it tells against the conditions.

Many psychologists today agree that—at least broadly speaking—our cognitive infrastructure is split into two systems. As Daniel Kahneman influentially put it, System 1 is the fast-processing system responsible for our unreflective, automatic judgments. System 2, in contrast, is the slow-processing, reflective system that makes it possible for us to engage in complex problem-solving, plan for the long term, consciously weigh reasons, and more. While System 2 is what allows us to engage in more sophisticated

cognitive activities, System 1 is unquestionably indispensable. Given that we encounter a superabundance of sensory information on a moment-to-moment basis and given that System 2 processing is laborious, we simply could not get on with our lives if our brains did not quickly process the vast majority of information with which we are regularly confronted (Kahneman, 2011). Thus, our brains regularly take "shortcuts" in service of the aim of fast and efficient processing. For example, we have a tendency to automatically and unconsciously compare and contrast new objects we encounter and objects with which we are already familiar.[12] This spares us the trouble of consciously and individually evaluating each new object we encounter.

Despite the practical and epistemic usefulness of System 1, judgments produced by it can also be epistemically problematic for the following reason: System 1 aims primarily at speed and efficiency and can produce false beliefs in service of those aims (Kahneman, 2011, especially pp. 105–107). In other words, while having efficiently formed beliefs might often overlap with having true beliefs—which would help explain why this tendency to generalize evolved—efficiency and truth can come apart. Thus, the judgments produced by System 1 are not always good from the epistemic point of view. Furthermore, the unreflective part of our brains responsible for making these judgments is not really in the business of determining when this is the case. Important to note for the purposes of our discussion is that this tendency to categorize, or associate, new with old is not limited to inanimate objects but applies to people as well (Gendler, 2011, 38–41). Thus, we regularly take shortcuts when making judgments about people, which can be epistemically (not to mention morally and politically) problematic.

To see the problem, we need to say a bit more about how this categorization works. First, because the purpose of System 1 shortcuts is to increase efficiency, the brain, when categorizing, highlights similarities among members of the same category as well as differences between members of different categories. In the context of social categorizing, this means that both in-group similarities and out-group differences are emphasized. Second, since our brains make sense of new people (as with all new objects) by comparing and contrasting them to ones with whom we already have experience, in an important sense, how we perceive the world is shaped by what we already believe

[12] This tendency to take shortcuts in judgments starts from an early age—infants compare and contrast new objects (e.g., a novel ball) with objects with which they have experience (e.g., familiar balls) and make assumptions about the new objects on this basis (see Baldwin et al., 1993). For a relevant discussion, see Leslie (2017).

and know (Banaji, 2002, especially p. 15102). When we identify someone as belonging to a certain group, we expect them to be like other members of that group, to behave in similar ways, and so forth. In short, to some degree we see what we expect to see. Third, the more familiar we are with certain categories—including what to associate with, and expect from, members of those categories—the more automatic and involuntary our judgments concerning those categories become.[13]

In addition, the ways in which we conceive of the groups with which we automatically associate people are likely socially informed. This means that racist, sexist, or otherwise discriminatory associations that have permeated the larger social context can distort how we conceive of the categories themselves. For example, the pervasive stereotype that Muslims are dangerous can distort the ways in which we come to think of, and treat, Muslim people; the pervasive stereotype that women are less self-assured or intelligent than men can distort the ways in which we come to think of, and treat, women; the pervasive stereotype that Black people are lazy or dangerous can come to distort the ways in which we think of, and treat, Black people; and so on.[14] And the more pervasive the stereotypes are, the more likely they are to influence how people conceive of certain groups. To take a concrete example, consider Keith Payne's (2001) studies, in which subjects identified guns faster, and misidentified tools as guns more often, when primed by non-White faces than they did when primed with White faces. The widely accepted explanation for this result is that the subjects were more likely to associate Blackness with danger than they were to associate Whiteness with danger.[15] Arguably the worst part about all of this is that once these associations become rigid and automatized, we may unconsciously stereotype people by associating them with certain groups—or as having certain characteristics—even when we consciously disavow these same stereotypes (Dovidio & Gaertner, 2000, 2004; Devine & Sharp, 2009; Gendler, 2011, pp. 41–44).

[13] For more discussion of all three of these points, see Gendler (2001, 39–40).

[14] Regrettably, the stereotypes that influence and infect the way we categorize people are numerous, and it is not difficult to think up other similar examples.

[15] Notably, when the subjects had more time to think before producing a judgment, the subjects' judgments were not as biased. However, Payne cautions us not to read too much into this shift in behavior as there is also evidence that people self-impose correcting measures depending on how they want to present themselves. In other words, the decrease in biased judgments when subjects were given more time was not necessarily due to their attempts to correct their biases—for example, because they were explicitly disavowing the stereotypes that were automatically triggered. Instead, it may have been due to their attempts to save face (Payne, 2001, p. 187). For more on attitudes and self-presentational strategies, see Fazio et al. (1995), Dunton and Fazio (1997), and Greenwald et al. (1998).

To sum up, our tendency both to pigeonhole people into categories with which we are already familiar and to expect certain things from them as a result is worrisome. Not only are the categories themselves often influenced by broader social norms and attitudes which can be misinformed, but also these norms and attitudes are often morally and politically harmful.

In addition to the worrisome evidence concerning System 1 reasoning, there is evidence that System 2 is not as rational (or reason-driven) as we might have assumed and that at least some types of judgments produced from this system are products of rationalization rather than proper reasoning. For instance, many people are strongly inclined to say that incest is wrong, even when it is consensual, it is done in secret (so nobody will find out and be appalled by it), protection is used, and no party to the incest will regret having done it in the future (Haidt et al., 2000, as cited in Haidt, 2001). In other words, the reasons typically cited for the wrongness of incest are not apposite. And yet, people tend to insist that it is immoral, even as they are empty-handed when pressed to come up with an explanation. According to Haidt's (2001) social intuitionist model of moral reasoning, what explains such cases is that people make moral judgments on the basis of intuition. In cases where someone is asked to explain why they made this or that moral judgment, they become "a lawyer trying to build a case rather than a judge searching for the truth" (Haidt, 2001, p. 814). His more general claim is that when people are presented with a moral problem, they respond intuitively and only later provide reasons, if at all. And, importantly, the reasons given are meant to accommodate a person's intuitions.[16]

In addition to what we have cited, there is a wealth of other research corroborating these and similar points. For instance, another type of bias (or family of biases) worth mentioning is confirmation bias, which can be summed up as the following pair of tendencies: (1) We tend to seek out, or take more seriously, evidence and argumentation that confirms what we already believe and (2) we tend to fail to seek out, or take seriously, evidence or argumentation that is in tension with what we believe (Nickerson, 1998; Kunda, 1999). These tendencies, coupled with the negative aspects of both System 1 and System 2 processing, should be worrying. It is bad enough that System 1 and System 2 can produce beliefs that are not well supported by the evidence but are rather the effects of biased thinking. What makes matters

[16] This, of course, is not to deny that people can determine that their moral intuitions are misguided. It is simply to suggest that, for many people and in general, our moral sensibilities may be guided less by reason and more by gut reaction than we recognize.

worse is that we will often continue holding these beliefs, even if they are not evidentially supported, since we tend to prioritize and preferentially treat evidence that confirms what we already believe.

To make matters even worse, there is also evidence suggesting that we do not have a reliable psychological capacity for telling the difference between beliefs formed on the basis of a given recognizable reason and beliefs formed via implicit bias. In particular, many people exhibit what's known as a *bias blind spot* (i.e., an inability to recognize when their own judgments are informed by biases). Thus, a general recognition that judgments are often formed (at least partially) in response to non-rational factors only gets us so far if everyone considers themselves an exception to the rule. Raising this very point, Pronin and Schmidt (2013) draw from a host of psychological studies which demonstrate, across a number of domains, the ignorance people have of their own biases. Especially germane to this discussion are the blind spots people have to their ideological, prejudicial, and self-interested biases. For instance, as Pronin and Schmidt discuss, people are prone to deny that their political or other ideological beliefs are formed as a result of partisan or ideological alignment, insisting that their beliefs were reached instead on the basis of sound reasoning (Cohen, 2003; Pronin et al., 2007; Robinson et al., 1995).[17] Likewise, people are prone to deny that their judgments are race- or gender-biased (Dovidio & Gaertner, 1991, 2004; Uhlmann & Cohen, 2005; Vivian & Berkowitz, 1992), as well as that their judgments or decisions have been influenced by monetary or social incentives (Dana & Loewenstein, 2003; Epley & Dunning, 2000; Heath, 1999; Miller & Ratner, 1998).

Perhaps the existence of the bias blind spot should come as no surprise as we have been aware for quite some time of the *Dunning-Kruger effect*, whereby people tend to both generally overestimate their abilities as well as rate their abilities as higher than those of their peers.[18] Assuming this tendency carries over to the problem of bias detection, then not only is it the case that we regularly overestimate our ability to detect when our thinking is biased but also the worse we are at detecting this, the better we think we are. Suffice it to say, the problem of pervasive bias is one we cannot ignore.

[17] This is true even though they are inclined to think their political or ideological opponent's views were largely shaped by their biases.

[18] The interesting, and especially worrying, aspect of this effect is that the less competency people have with respect to a certain skill, the more drastic their overestimation of their own capabilities. For more on this, see Kruger and Dunning (1999). It is also worth noting that not everyone suffers from a tendency to overestimate their abilities. Some people have the opposite sort of problem—namely, they regularly underestimate their abilities, rating themselves as less capable than they in fact are.

Of course, despite all of this alarming data, recall that not all associative judgments that could be labeled as biased judgments are necessarily "bad," epistemically or otherwise. For instance, quick associative judgments produced by System 1 processing are often prudentially and epistemically useful—after all, if we did not take shortcuts in reasoning, it would be nearly impossible to get on with our lives. With this in mind, we can make an intuitive distinction between cases of "good" bias—cases where biases and unconscious cognitive heuristics lead us to make better and more justified judgments—and what we'll call cases of "bad" bias:

> *Bad Bias*: S forms a belief that *p* on account of bad bias just when S's believing that *p* is the result of reason-independent factors that should, were S aware of it, cause them to not so believe.[19]

Let's say a *bad-bias situation* is a situation in which one's belief is formed because of bad bias. By *reason-independent* factors we simply mean those factors—whatever they turn out to be as the science unfolds—which the above-mentioned research has labeled "implicit bias," "confirmation bias," "cultural conditioning," or simply prejudice. At minimum, such evidence points to the extreme difficulty of ever meeting the criteria of response and reflection.[20] This, in turn, calls into question the likelihood of ever meeting the norm of reasonableness.

Against reflection, this evidence suggests that beliefs are often not the causal products of reasoning—where reasoning involves being responsive to facts and making proper inferences from those facts—but rather the products of bad bias. This happens when, for instance, we take the testimony or intellectual contributions of citizens who are perceived as belonging to stigmatized groups less seriously than the testimony or contributions of non-stigmatized citizens.[21] And if we are responsive not just to reasons but often also to reason-independent factors, then it becomes unclear how well this condition can be met. This is especially true assuming that—as the studies

[19] More on the "should" here: We certainly want to make room for the possibility that—given that none of us is an ideal rational agent—a person might come to learn that their act of believing that *p* is the result of reason-independent factors. And, though from the epistemic standpoint this ought to cause them to cease believing that *p*, their belief that *p* is in no way affected (perhaps because they are dogmatic about *p*). Such an agent is not doing what they epistemically should be doing.

[20] Defenses of both anti-response and anti-reflection can also be found in Kornblith (2012).

[21] We are currently interested in the stigmatization of groups that have historically been oppressed or marginalized in morally problematic ways—for instance, people of color, women, the LGBTQI+ community, and more.

on moral reasoning and confirmation bias indicate—we have tendencies that make us overlook or discount reasons when they conflict with what we already take to be true.

A different, and arguably more troubling, problem emerges in relation to response: Even if people can recognize reasons as such, they may still very well fail to see those reasons as applicable to them as a result of bias blind spot. For instance, although someone might recognize and acknowledge the problems implicit biases pose for a healthy democracy, they might think that they tend to be less biased in their thinking than the average citizen. So—thinking that their beliefs are not affected by stereotypes or that they do not suffer from confirmation bias—they very well may be unresponsive to certain reasons that are presented to them. And so, the difficulty in meeting the norm of reasonableness has not disappeared. Moreover, assuming the Dunning-Kruger effect carries over to bias detection, not only are most people likely to count themselves as some of the lucky few whose judgments are not largely biased but also the worse people are at determining when their judgments are formed as a result of bad bias, the more likely they are to think that their thinking is largely unbiased.

A Skeptical Argument

Should all of this evidence, taken together, encourage skepticism about our ability to be reflective and responsive to reasons? As a matter of fact, similar considerations have encouraged skepticism about reason throughout philosophical history.[22] Consider what Hume had to say:

> When I reflect on the natural fallibility of my judgment, I have less confidence in my opinions than when I only consider the objects concerning which I reason; and when I proceed still farther, to turn the scrutiny against every successive estimation I make of my faculties, all the rules of logic

[22] We agree with Saul (2013) that this kind of skeptical argument can be seen as more worrisome than most. As she notes, most skeptical problems are either live (in the sense that they cannot be ignored) or global (in the sense that they have wide range—e.g., all of perceptual experience). But they are not both. Certainly, this skeptical problem is live. Arguably, it is global, too. For—though it does not force us to call into question all of our perceptual beliefs, as most traditional forms of skepticism do—it does force us to come face to face with the possibility that a great number of our doxastic states are influenced by bad bias. In our view, considering the effects of bad bias in relation to the norm of reasonableness provides a stark illustration of this fact.

> require a continual diminution, and at last a total extinction of belief and evidence. (Hume, 1739/1896, Book I, Part IV, Section 1, p. 183)

Abstracting from the details of his particular argumentative agenda, Hume's worry is that simply reflecting on his own fallibility undermines his faith in his reasoning. More recently, Kornblith (1999, 2012) and Saul (2013) have appealed to similar contemporary psychological research to motivate skeptical worries. We can, in fact, collect these worries and shape them into the form of a traditional skeptical argument:

Bad-Bias Argument

1. If I am justified in believing *p* on the basis of a given, recognizable reason R,[23] then I am justified in believing I am not in a bad-bias situation with respect to my belief that *p*.
2. I am not justified in believing I am not in a bad-bias situation with respect to my belief that *p*.
3. Therefore, I am not justified in believing *p* on the basis of R.

Let's consider this argument. In defense of the first premise, if someone were justified in believing *p* on the basis of a given, recognizable reason R, then they would be justified in believing that their belief that *p* is not formed on the basis of reason-independent factors. That is, they would be justified in believing that they are not in a bad-bias situation. However, in the previous section we considered evidence which points to the strong possibility that we are often not justified in believing that we are not in a bad-bias situation.

To repeat, against reflection, there is quite a bit of evidence suggesting that beliefs are frequently formed as the result of bad bias. Additionally, against response, the evidence suggests that we are bad at detecting when our beliefs are formed as the result of bad bias. This evidence, taken together, gives us premise 2: I am not justified in believing I am not in a bad-bias situation. After all, if my beliefs are often formed as the result of bad bias, and if I truly am lousy at detecting when beliefs are formed in this way versus when they are not, then I cannot—at least with respect to certain beliefs (this caveat will be explored below; see "Mitigating Responses")—rule out the possibility that I am in a bad-bias situation. This, of course, brings us to the conclusion,

[23] By "recognizable reason," we have in mind the following: Person A gives a recognizable reason R to person B only if B would recognize that R is a reason were B to reflect on it and reason consistently with B's epistemic principles. See Lynch (2012) for further discussion.

which is that I am not justified in believing *p* on the basis of a given, recognizable reason R. But if this is true, then it is hard to see how the norm of reasonableness could ever be met. So much the worse, one might think, for public discourse.

Objections to the Argument

Like any skeptical argument, the bad-bias argument can be attacked in a variety of ways. While we can hardly attempt a comprehensive coverage of possible responses here, we'll start by noting that they can come in two basic kinds. What we'll call *unraveling responses* are direct objections to the soundness of the skeptical argument. *Mitigating responses*, on the other hand, accept (some version of) the argument but attempt to blunt its force. We will consider responses of both kinds, beginning with three kinds of unraveling responses.

Unraveling Objections

First, one might resist the argument by rejecting the closure principle, upon which the argument implicitly relies. While different variations of the closure principle exist, we cite Pritchard's (2016) closure$_{RK}$ principle:

> If S has rationally grounded knowledge that *p*, and S competently deduces from *p* that *q*, thereby forming a belief that *q* on this basis while retaining her rationally grounded knowledge that *p*, then S has rationally grounded knowledge that *q*. (p. 23)[24]

Rejecting the closure principle would allow one to reject Premise 1, and this is certainly one way to resist the undesirable conclusion. However, given that the closure principle has such strong intuitive plausibility, one would presumably need reasons for rejecting it other than that it allows us to reject the particular skeptical argument under consideration. One such reason might be that rejecting it would allow us to avoid skeptical arguments altogether.

[24] For reasons to prefer this version of the closure principle to others, see Pritchard (2016, especially pp. 11–25).

However, this has the air of throwing the baby out with the bathwater and does not seem to be the most promising way of resisting the argument.[25]

A more plausible objection rests on the idea that the argument is self-undermining. After all, it seems that if the argument is sound, then I cannot be justified in believing its conclusion on the basis of hearing the argument—since (a) hearing the argument means being given a reason for the conclusion and (b) according to the argument itself, that means my belief in the conclusion can only be justified if I am justified in believing I am not in a bad-bias situation. But the argument contends that I am not so justified, and thus I am not justified in believing its conclusion when given the argument as a reason to believe that conclusion.

Hilary Kornblith (1999)—who, as noted, is one of the few philosophers to explicitly consider something like the skeptical argument outlined—has suggested a response to this objection on behalf of the skeptic. Skeptical arguments, Kornblith argues, are meant to leave us with what we might call the presumption of epistemic guilt. They tell us our beliefs are guilty of epistemic crimes. If this is right, then the skeptical conclusion in question should really be read as follows: Beliefs formed on the basis of recognizable given reasons are guilty until proven (epistemically) innocent (Kornblith, 1999, p. 189).

The point of Kornblith's maneuver is roughly this: While the psychological evidence suggests we should assume that any belief we form on the basis of a given reason is infected with bad bias, some beliefs formed on reasons (like maybe this one—i.e., the very belief that some beliefs are infected with bad bias) might still turn out to be justified. That is, we might examine the belief and later find further evidence to think that it is true. The same goes for the skeptical argument. We may not be justified in believing in it—for instance, when first encountering it in a public setting (given the possibility of bad bias). But further investigation may clear it of guilt and set it free.

Continuing Kornblith's line of thought, perhaps every belief we form via an exchange of reasons should be presumed to be guilty, but perhaps many of them can also be saved. Indeed, this is sometimes the very approach we seem to take in ordinary life, where people are often willing enough to grant

[25] One might also protest that the consequent of the first premise of the argument places an unnecessary higher-level constraint on the justification of beliefs formed via recognizable reasons. But it is unclear that such a constraint is unnecessary—given that the context of public discourse is explicitly discursive and reflective—nor is it clear that the constraint is viscous, for the consequent only requires that I be justified in believing my belief that *p* is not formed on the basis of bad bias, not that I consciously do so at the moment I form my belief that *p*.

that biases negatively impact one's ability to reason well. This point has merit, but we are not convinced that Kornblith's strategy alone saves the argument from the charge that it is self-undermining. That's because to prove that a given belief is epistemically innocent of bad bias, one must presumably engage in reason-giving. Arguments in the epistemic court for innocence are, in this sense, the same as arguments for guilt. They all trade in reasons. So any attempt to show (in epistemic court, as it were) that one's belief in the conclusion of the skeptical argument is itself free from bad bias will itself involve appeals to reasons. And whatever those reasons may be, if one forms a belief—such as the belief that the conclusion of the skeptical argument is free from bad bias—on the basis of those reasons, that belief, according to the skeptical argument, will be unjustified. Hence, if the argument is sound, then any attempt to show—by providing a reason—that its conclusion is free from bad bias will itself be unjustified.

There is, however, another strategy the skeptic might use to defend the argument. They might claim that the argument has a restricted scope and that the conclusion of the argument does not fall within that scope. They might claim, for example, that the argument only applies to those beliefs that can be affected by bad bias, for example, first-order beliefs concerning social, political, religious, and moral matters. If so, then the conclusion—being about whether a given belief is justified—may not be within that scope. Otherwise put, the skeptic may claim that the bad-bias argument does not apply to epistemological and logical matters. Thus, the argument does not fall on its own sword.

Unfortunately for the skeptic, that sword may be sharper than this reply allows; for, as we have indicated, the effects of bad bias often arise due to *facts about who is giving you the relevant reason, not the content of the reason itself*. It is the testifier, not the testimony, that frequently matters. If so, then the possibility of bad bias remains even when the content in question is of a perfectly uncontroversial type. Thus, for example, bad bias might result in a mathematician's work not being taken seriously even when that work is technically sound, simply because of their gender, race, or both. Given this, the skeptic may find it difficult to limit the scope of the argument, and hence avoid the charge of being self-undermining as a result.

So, is the bad-bias argument self-undermining or not? Yes and no. Consciously given as a reason (whether to myself or others) to be skeptical about one's beliefs, it may well be. But its soundness and the truth of its conclusion do not hinge on whether it is given as a reason. If it is sound, then

my beliefs based on given recognizable reasons—or at least those possibly effected by bad bias—are not justified by those reasons. This is true whether or not I or anyone else ever recognizes that fact.[26] What this suggests, at the very least, is that the argument is no more easily defeated than other skeptical arguments.

Mitigating Responses

We now turn to a different kind of response to the bad-bias argument: responses that seek not to refute the argument totally but rather to blunt its force.

In one sense, we've already encountered such a response—namely, that the argument is limited in its scope. In the previous section, we imagined this as a possible way the skeptic could respond to the charge that the argument is self-undermining. However, it can also be seen as a mitigating response to the argument. Not only can it be seen as a way for the skeptic to save the argument but also—assuming the argument can be saved—this response serves to weaken the force of the argument. As noted, however, the virtue of this response rests in part on ignoring the fact that bad bias is often the result of who is giving the reason in question, not the content of that reason. As such, the response may have limited value.

Alternatively, one might argue that the bad-bias argument shows, at best, that I am not justified in believing a proposition based on a given recognizable reason. That does not entail that the belief is unjustified. My beliefs in certain mathematical and logical propositions, for example, might be justified by a priori intuition or because they are self-evident and not on the basis of any recognizable reason.[27] Consider, too, that small children and animals arguably have justified perceptual beliefs even though they are incapable of articulating, let alone recognizing, why their beliefs are justified. While this response does allow us to gain some ground, it only gets us so far, for—according to response and reflection—often what we care about in public discourse is being able to form justified beliefs on the basis of recognizable reasons. Moreover, we ought to be able to recognize when this is the case. We

[26] Of course, whether it is sound depends on other factors as well, including some we mentioned at the beginning of this section, such as the closure$_{RK}$ principle.

[27] Whether one accepts this point and the ones that follow in this paragraph rests, of course, on how one construes concepts like *self-evidence*, *the a priori*, and *propositional justification*.

aim to exercise in the public square an ability that sets us apart from small children and animals—namely, our ability to engage in the practice of giving and asking for reasons, which presumably requires the ability to discriminate good reasons from bad.

The third—and, we think, most compelling—attempt to lessen the force of the argument is to say that Premise 1 is not obviously supported by the evidence cited previously (see "Against Reflection and Response"). After all, it is not as if having our beliefs or judgments affected by biases and other epistemically suspect non-rational factors is an all-or-nothing kind of matter. Instead, how much one is affected by non-rational factors when forming beliefs is presumably a matter of degree—often, it is neither solely objective reasons nor reason-independent factors that shape one's beliefs but rather some combination of the two. On account of this, one might argue that a more nuanced, and perhaps less threatening, skeptical argument is actually supported by the evidence we've considered. Here is one way to run such an argument (we emphasize it is not the only way):

Modified Bad-Bias Argument

1. To the extent that I am fully justified in believing *p* on the basis of a given, recognizable reason R, then to that same extent I am fully justified in believing I am not in a bad-bias situation.
2. I am not fully justified in believing I am not in a bad-bias situation.
3. Therefore, I am not fully justified in believing *p* on the basis of a given, recognizable reason R.

While understanding bad-bias "infection" as coming in degrees does not allow us to completely evade the skeptical worry, it does drive a wedge between being in a bad-bias situation and being unjustified in believing *p* on the basis of a given, recognizable reason R. This is because one could be partially in a bad-bias situation—in the sense that biases are negatively affecting one's judgments to some, but less than full, extent—but also not completely unresponsive to reason. In this type of situation, to the extent that one's belief formation is negatively affected by reason-independent factors, then to that same extent one is not justified in believing that one's belief that *p* was formed on the basis of a given, recognizable reason R. Notice, however, that this leaves room for one being partially justified in believing that one's belief that *p* was formed on the basis of a given, recognizable reason R.

One problem with this response is that it appears to only push the problem back, for now we are left with the burden of determining not just when our judgments are infected by bias but also the extent to which they are so. And if we are bad at determining when our thinking is biased, we are likely also bad at determining how much our thinking is biased. Yet there may be a silver lining: Even if we cannot be confident of our ability to determine when, or by how much, our judgments are formed as the result of bad bias, the point does underline an important possibility, namely that there may be indirect ways to lower the degree to which bad bias infects our thinking. Such a possibility is crucial because one thing seems clear: Even if the bad-bias argument is unsound (in either version presented), there is significant inductive evidence that the response and reflection conditions are extremely difficult to meet. And that fact alone is worrying enough for those like us concerned with the norm of reasonableness.

Intellectual Humility and the Threat of Bad Bias

We want to conclude by suggesting—if only in a rough, initial fashion—two ways we may yet be able to counter the threat posed by bad bias to the norm of reasonableness. Our touchstone is that there may be indirect ways to lessen the impact of bad bias on our thinking, and therefore lessen the skeptical threat posed by it.

Somewhat paradoxically, we begin with a simple fact that is nonetheless easy to overlook: There is arguably an epistemic benefit from thinking about the possibility that one's beliefs formed by reasons could be affected by bad bias. The benefit we have in mind does not directly impact the positive epistemic status of our beliefs; rather, it can cause us to have a particular epistemic attitude, the having of which can indirectly affect our beliefs and the reliability of the processes that produce those beliefs. The attitude we have in mind—and the focus of one of the conditions we have listed—is what is sometimes called "epistemic" or "intellectual" humility. While the nature of this attitude—and even whether it is an attitude rather than a trait—is a matter of debate, we take it that two significant marks of having the attitude concern the following:

1. Owning one's intellectual limitations (Whitcomb et al., 2015; Tanesini, 2016; Lynch, 2018b)

2. Being willing and able to learn from other people's testimony and experience—that is, being willing and able to revise one's beliefs and attitudes in light of evidence supplied by others (Lynch, 2018a, 2018b).[28]

As we see it, intellectual humility is both a self-regarding and an other-regarding attitude. It concerns seeing oneself as a limited cognitive being, capable of being affected by bias and prejudice. But really being humble in this way also means being willing and able to listen to others—that is, to think that one alone can't know it all.

Our present point is that one benefit of considering (either version of) the bad-bias argument and the associated evidence against response and reflection is that doing so gives us good reason to be more intellectually humble, for intellectual humility arguably counteracts the effects of bad bias. The same goes for cognitive empathy, which is in an important sense intimately caught up with intellectual humility. This is because—in order to see one's worldview as capable of improvement and to be willing and able to revise one's beliefs in light of the evidence and testimony provided by others—one arguably has to have the imaginative capacity to see where others are coming from. Thus, considering the possibility of bad bias gives us good reason both to try to see where others are coming from and to own our limitations to a greater degree (by recognizing that we can be affected by such bias). We may then, in turn, be more willing and able to revise our beliefs in response to reasons.

Of course, being more intellectually humble and cognitively empathetic does not mean that any arbitrary belief we may have will automatically be free from bad bias. Nonetheless, there is evidence to suggest that in the long run pointing out instances of bad bias can cause people to be more receptive to alternative points of view.[29] If so, then the facts about bad bias we have cited may well encourage cognitive empathy and intellectual humility in those who encounter them. Moreover, in this context of our discussion

[28] For further discussion of intellectual humility, see Hazlett (2012), Spiegel (2012), Christen et al. (2014), Whitcomb et al. (2015), Church (2016), Kidd (2016), and Tanesini (2016).

[29] Consider, for instance, the public's response to the infamous (since now widely recognized as racist) 1988 Willie Horton advertisement launched by the George H. W. Bush campaign against Michael Dukakis. Dukakis had been ahead in opinion polls until the ad was released, at which point he lost favor to Bush. However, when Jesse Jackson called the ad out for being racist, support for Bush began to decline. For more on this—including a discussion of why this correlation between Jackson's remarks and decreased support for Bush is plausibly more than mere correlation—see Mendelberg (2001, especially Chapters 5–8) and Saul (2018, especially pp. 10–12).

of democracy and the norm of reasonableness, it is worth noting that even if the consideration of bad bias only increases cognitive empathy and intellectual humility in some parts of the population, that alone may make the ideal—considered as a general norm on discourse conducted throughout that population—more tractable. In other words, not everyone has to be that cognitively empathetic and humble; increasing these attitudes in some people may increase the overall reasonability of public discourse.

This brings us to a second way we might indirectly lower the threat posed by bad bias to the norm of reasonableness. In addition to (a) focusing on our beliefs themselves and determining whether they are the products of bad bias or (b) trying individually to be more intellectually humble and empathetic, we can (c) focus our energy on positively transforming social practices. That is, we can promote the realization of epistemic equality by encouraging social and institutional practices that have the effect of making their participants more intellectually humble.[30] Consider, for example, the following very different norms or rules that can be incorporated into certain social practices:

- Requiring police review boards to be racially and ethnically diverse and to include community members
- Obligating journalists to provide independent sources for claims made as fact in a story
- Requiring surgeons to utilize checklists prior to surgery (Gawande, 2009)

As noted, each of these suggested rules is very different in context and content. Yet each can be seen as encouraging those participating in the relevant practices to have either the first (1) or the second (2) aspect of intellectual humility we have highlighted. And they do so for a simple reason: They encourage the participants to see themselves as limited beings who don't know it all. A diverse community review board can make the possibility of implicit bias salient to the practice of policing in that community. This in turn allows for contexts where specific acts of policing can be called out for being based on beliefs which are infected with bad bias; it also encourages police to listen to the concerns of community members. Likewise, the process of double-checking and duplicating sources can make errors due to bad bias—on the part of either a source or a reporter—more salient and obvious, and it can

[30] We see the following point as being more applicable to intellectual humility, though perhaps it is applicable to cognitive empathy as well.

encourage reporters to seek out those who may contradict an initial source. The same goes for surgical checklists. A checklist—whether that of a pilot or a surgeon—has an explicitly epistemic point: It literally acts as a check for unnoticed error. By following the practice of consulting a checklist prior to, for example, amputating a limb, the surgeon engages in an activity meant to account for the fact that they have cognitive limitations and can make mistakes.

More can be said about each of these suggestions. Our present point is that one way to combat the negative effects of bad bias is to support institutional norms and practices that encourage intellectual humility—that is, norms and practices that encourage an attitude which acknowledges the negative effects of bad bias and motivates people to counter such effects.

While the focus in the first half of the chapter was on response and reflection, the discussion in this section has brought us back full circle, for if we are right, the importance of the other conditions to the norm of reasonableness now becomes apparent. Here is why: While bias-related doubt puts pressure on response and reflection, trying to promote the realization of the other conditions—intellectual humility, epistemic equality, and cognitive empathy—might go some way to relieving the pressure. In other words, bias-related doubt initially made us worry that we can't meet the norm of reasonableness, given that the prevalence of biases makes it unclear whether we can meet the two necessary conditions of response and reflection. However, promoting the realization of the other conditions—intellectual humility, epistemic equality, and cognitive empathy—might put the plausibility of meeting the norm back on the table. After all, the more institutional equality there is and the more that citizens embody cognitive empathy and intellectual humility, the less likely it is, arguably, that our judgments will be affected by bad bias. Or, at least, our judgments will be affected by bad biases to a lesser extent.

Conclusion

The pervasiveness of bad bias is a clear threat to meaningful and epistemically useful public discourse, for it raises the possibility that the beliefs we form on the basis of such discourse are often the result of bias, not reasons. This in turn raises the possibility that democracy—understood as a space of reasons—is an unfulfillable ideal. This threat must be taken seriously, and we have attempted to engage with it seriously in this chapter. However, this threat does not signal that all is lost. What it does signal, or so we have argued,

is that if we wish public discourse to be more reasonable, we need to increase our efforts to develop intellectual humility, cognitive empathy, and epistemic equality, both individually and through our social practices.

References

Anderson, E. (2006). Epistemology of democracy. *Episteme*, *3*(1–2), 8–22.

Baldwin, D. A., Markman, E. M., & Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development*, *64*(3), 711–728.

Banaji, M. R. (2002). Social psychology of stereotypes. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 15100–15104). Pergamon.

Christen, M., Alfano, M., & Robinson, B. (2014). The Semantic Neighborhood of Intellectual Humility. In A. Herzig & E. Lorini (Eds.), *Proceedings of the European Conference on Social Intelligence* (Vol. 1283, pp. 40–49). CEUR-WS.org.

Church, I. M. (2016). The doxastic account of intellectual humility. *Logos and Episteme*, *7*(4), 413–433.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology*, *85*, 808–822.

Dana, J., & Loewenstein, G. (2003). A social science perspective on gifts to physicians from industry. *Journal of the American Medical Association*, *290*, 252–255.

Darwall, S. L. (2013). *Honor, history and relationship: Essays in second-personal ethics II*. Oxford University Press.

Devine, P. G., & Sharp, L. B. (2009). Automatic and controlled processes in stereotyping and prejudice. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 61–82). Psychology Press.

Dovidio, J. F., & Gaertner, S. L. (1991). Changes in the expression and assessment of racial prejudice. In H. J. Knopke, R. J. Norrell, & R. W. Rogers (Eds.), *Opening doors: Perspectives of race relations in contemporary America* (pp. 119–148). University of Alabama Press.

Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*(4), 315–319.

Dovidio, J. F., & Gaertner, S. L. (2004). Averse racism. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 36, pp. 1–52). Academic Press.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*, 316–326.

Epley, N., & Dunning, D. (2000). Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology*, *79*, 861–875.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*, 1013–1027.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Gaus, G. (2012). *The order of public reason: A theory of freedom and morality in a diverse and bounded world*. Cambridge University Press.

Gawande, A. (2009). *The checklist manifesto: How to get things right*. Henry Holt and Company.

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, *156*, 33–63.

Goldman, A., & Whitcomb, D. (Eds.). (2011). *Social epistemology: Essential readings*. Oxford University Press.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Habermas, J. (1985). *The theory of communicative action: Reason and the rationalization of society* (T, McCarthy, Trans.; Vol. 1). Beacon Press.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Haidt, J., Björklund, F., & Murphy, S. (2000). *Moral dumbfounding: When intuition finds no reason* [Unpublished manuscript]. University of Virginia.

Hazlett, A. (2012). Higher-order epistemic attitudes and intellectual humility. *Episteme*, *9*(3), 205–223.

Heath, C. (1999). On the social psychology of agency relationships: Lay theories of motivation overemphasize extrinsic incentives. *Organizational Behavior and Human Decision Processes*, *78*, 25–62.

Hume, D. (1896). *A treatise of human nature*. Edited by L. A. Selby-Bigge. Clarendon Press. (Original work published 1739)

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

Kidd, I. J. (2016). Intellectual humility, confidence, and argumentation. *Topoi*, *35*(2), 395–402.

Kornblith, H. (1999). Distrusting reason. *Midwest Studies in Philosophy*, *23*, 181–196.

Kornblith, H. (2012). *On reflection*. Oxford University Press.

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.

Kunda, Z. (1999). *Social cognition: Making sense of people*. MIT Press.

Leslie, S.-J. (2017). The original sin of cognition: Fear, prejudice, and generalization. *The Journal of Philosophy*, *118*(8), 393–421.

Lynch, M. P. (2012). *In praise of reason: Why rationality matters for democracy*. MIT Press.

Lynch, M. P. (2016). After the spade turns: Disagreement, first principles and epistemic contractarianism. *International Journal for the Study of Skepticism*, 6, 248–259.

Lynch, M. P. (2018a). Epistemic arrogance and the value of political dissent. In C. R. Johnson (Eds.), *Voicing dissent: The ethics and epistemology of making disagreement public* (pp. 129–139). Routledge.

Lynch, M. P. (2018b). Arrogance, truth, and public discourse. *Episteme*, *15*(3), 283–296.

Medina, J. (2012). Hermeneutical injustice and polyphonic contextualism: Social silences and shared hermeneutical responsibilities. *Social Epistemology*, *26*(2), 201–220.

Medina, J. (2018). Epistemic injustice and epistemologies of ignorance. In P. C. Taylor, L. M. Alcoff, & L. Anderson (Eds.), *Routledge companion to the philosophy of race* (pp. 247–260). Routledge.

Mendelberg, T. (2001). *The race card: Campaign strategy, implicit messages, and the norm of equality*. Princeton University Press.

Miller, D. T., & Ratner, R. K. (1998). The disparity between the actual and assumed power of self-interest. *Journal of Personality and Social Psychology*, *74*, 53–62.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220.

Paul, L. A. (2014). *Transformative experience*. Oxford University Press.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181–192.

Pritchard, D. (2016). *Epistemic angst*. Princeton University Press.

Pronin, E., Berger, J. A., & Molouki, S. (2007). Alone in a crowd of sheep: Asymmetric perceptions of conformity and their roots in an introspection illusion. *Journal of Personality and Social Psychology*, 92, 585–595.

Pronin, E., & Schmidt, K. (2013). Claims and denials of bias and their implications for policy. In E. Shafir (Ed.), *The behavioral foundations of public policy* (pp. 195–216). Princeton University Press.

Rawls, J. (1971). *A theory of justice*. Belknap Press.

Robinson, R. J., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: "Naïve realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology*, *68*, 404–417.

Saul, J. (2013). Skepticism and implicit bias. *Disputatio*, *5*(37), 243–263.

Saul, J. (2018). Dogwhistles, Political Manipulation and Philosophy of Language. In D. Fogal, D. W. Harris, & M. Moss (Eds.), *New Work on Speech Acts* (pp. 360–383). Oxford University Press.

Spiegel, J. S. (2012). Open-mindedness and intellectual humility. *School Field*, *10*(1), 27–38.

Talisse, R. (2009). *Democracy and moral conflict*. Cambridge University Press.

Tanesini, A. (2016). Intellectual humility as an attitude. *Philosophy and Phenomenological Research*, *96*(2), 399–420.

Uhlmann, E. L., & Cohen, G. L. (2007). "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Process*, *104*, 207–223.

Vivian, J. E., & Berkowitz, N. H. (1992). Anticipated bias from an outgroup: An attributional analysis. *European Journal of Social Psychology*, *22*, 415–424.

Whitcomb, D., Battaly, H., Baehr, J., & Howard-Snyder, D. (2015). Intellectual humility: Owning our limitations. *Philosophy and Phenomenological Research*, *94*(3), 509–539.

5

Rationalization, Creativity, and Imaginative Resistance

Jason D'Cruz

The Point of Rationalization

My explanatory target is rationalization in the sense of spurious self-justification.[1] I argue that we should model episodes of rationalization as fictional narratives in which the rationalizer plays the starring role. I bring this model of "rationalization as pretense" into communication with the philosophical literature on imaginative engagement in fiction and *imaginative resistance*, the phenomenon whereby an otherwise competent imaginer experiences difficulty in taking part in an imaginative activity (Gendler & Liao, 2016). While rationalization is most often modeled as a motivationally biased, and thereby faulty, form of deliberative belief formation, it is easy to overlook that successful rationalization represents a kind of cognitive and creative achievement. The achievement consists in crafting a narrative whose rehearsal has the effect of eliciting desirable emotions and mitigating feelings of guilt or shame. In contrast to the intellectual activities of inquiry or doxastic deliberation, rationalization is an inherently creative undertaking. *Rationalizing* is the process of generating and rehearsing narratives that have the credible appearance of genuine deliberation and inquiry but whose narrative arc aims at exculpation or self-justification.

The empirical finding that inveterate deceivers tend to be highly creative individuals (Gino & Ariely, 2012) fits well with a picture of rationalization as an inherently inventive activity rather than a faulty investigative one. As Ariely puts it, "If the key to our dishonesty is our ability to think of ourselves as honest and moral people while at the same time benefitting from cheating,

[1] This use of the term should be kept separate from Davidson's (1963) technical use of the term *rationalization* to mean "providing reasons to explain an agent's action."

Jason D'Cruz, *Rationalization, Creativity, and Imaginative Resistance* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press.
DOI: 10.1093/oso/9780197636916.003.0006

creativity can help us tell better stories—stories that allow us to be even more dishonest but still think of ourselves as wonderfully honest people" (2012, p. 188).

Rationalizers pursue two aims in tandem. The first is to craft a story that has the right kind of narrative arc, most often one that explicitly or implicitly justifies suspect behavior. The second aim is for the story to be plausible enough to sustain the suspension of disbelief. If things go right, the rationalizer feels little skepticism toward the story they concoct, despite its inconsistencies. This allows them to respond to the imagined content with strongly felt emotion (Holland, 2008). In this way successful rationalizers mitigate feelings of guilt or shame and cultivate feelings of indignation or self-righteousness. Successful rationalization is emotion regulation working well rather than reasoning working poorly. (Of course, this is not to deny that rationalization brings with it serious moral and practical liabilities.)

Rationalization is typically past-directed, as suggested by the phrase *post-hoc rationalization*. In explaining away our culpability or irrationality, rationalization can make us feel better about the things we have done. For example, we may generate a "just so" story about why a particular action does not count against our good character or good judgment.[2] *Everyone else is cheating too, so my cheating just levels the playing field*. Or, *The reason I missed the deadline is that I'm such a perfectionist about my work*. But rationalization can also be anticipatory, helping to clear away hurdles of caution and conscience when we consider prudentially or morally dubious courses of action. Anticipatory rationalization has implications for practical reasoning. The *hoc* in *post hoc* may refer to decisions or inclinations relating to future actions rather than the actions themselves. Recent work by Shalvi et al. (2015) suggests that self-justifications that take place temporally prior to ethical violations enable people to continue to feel good about themselves while doing things they know or suspect to be immoral. "Previolation" rationalizations serve to defuse the anticipated threat to the moral self.

Effective rationalization is useful, but it is not easy. Rationalizers do not simply arrive at their conclusions "at will" in the way you could, for example, simply decide to imagine a brown bear in the corner of the room. Rather, the rationalizer must work skillfully with the available evidence. But since rationalizers do not aim at truth or understanding, they have a degree of

[2] Preservation of self-concept is central to Elliot Aronson's (1969) influential reformulation of Festinger's cognitive dissonance theory.

flexibility. Rationalizers arrive at their conclusions via a process that is most aptly described as *continuously partially constrained* by their appraisal of the evidence. A major advantage of the model I propose is that it explains how such a relationship to the evidence is possible. On this model, rationalizers make as if they are guided in deliberation by the norm of truth (accept *p* only if *p* is true) but are in fact is guided by the norm of plausibility (accept *p* only if *p* appears true from a particular vantage). What makes rationalization effective in mitigating feelings of guilt and shame and in cultivating feelings of self-righteousness and confidence are the systematic connections between imagination and emotion. Attending to these connections allows us to better understand the factors that either catalyze or forestall rationalization. To characterize these factors more precisely, I draw on the philosophical and psychological literature on imaginative engagement with fictional narratives.

I begin with a presentation of the central features of the model of rationalization-as-pretense.

A Model of Rationalization and an Illustration

The model of rationalization-as-pretense is designed to reconcile the following three constraints[3]:

1. *Deliberative exclusivity*: A thinker cannot in full consciousness decide whether to believe *p* in a way that issues directly in forming a belief by adducing anything other than considerations that they regard as relevant to the truth of *p* (D'Cruz, 2014, 2015).
2. *Non-naïveté*: Rationalizers know or suspect that the considerations they adduce are not sufficient to establish the conclusions they reach.
3. *Deliberative weighing*: The considerations adduced in the process of rationalization play an essential role in the deliberative formation of the conclusion.

The first thesis, deliberative exclusivity, is widely endorsed by commentators in the literature on the "aim of belief"—a debate about whether believing a proposition carries with it a commitment or teleological directedness toward the truth of that proposition. This includes normativists (Boghossian,

[3] See D'Cruz (2014, 2015).

2003; Engel, 2013; Shah & Velleman, 2005; Wedgwood, 2013) (who hold that it is a conceptually constitutive normative feature of beliefs that they ought to be true), teleologists (McHugh, 2012; Steglich-Petersen, 2009; Vellman, 2000) (who hold that belief aims at truth in the psychological sense that beliefs are intended by agents or regulated by subpersonal mechanisms to be true), and skeptics (Glüer & Wikforss, 2009; Hazlett, 2013; Owens (2000, 2003)) (who hold that various formulations of the aim thesis are false or platitudinous). From the perspective of first personal doxastic deliberation (deliberation about what to believe), only considerations that appear to the subject as relevant to the truth of the proposition being considered can have an influence on the deliberative outcome. Indeed, some of participants in the debate on the aim of belief take it as an important desideratum that their theories account for this aspect of the phenomenology of doxastic deliberation. From the perspective of first personal doxastic deliberation, whether to believe a proposition is exclusively a matter of whether the proposition is true: Questions of a person's practical aims or moral commitments don't play an explicit role. It is noteworthy that even pragmatists (who maintain that there are non-evidential reasons for belief) deny that we ever explicitly evaluate the rationality of our beliefs in terms of how well they promote our goals: "Offering you a million dollars to believe that the earth is flat may convince you that you have a good economic reason to believe the proposition, but in itself it won't be enough to persuade you that the earth is really flat" (Foley, 1993, p. 16).

The second thesis, non-naïveté, is supported by the responsiveness that rationalizers display to the insufficiency of the considerations they adduce in establishing the conclusions they reach, as well as by their abandonment of their rationalizations when their incentives change. Evidence that rationalizers suspect that the considerations they adduce fail to establish the conclusions they reach comes from two sources: strategic avoidance of evidence and cases of "trumped incentive." Eric Funkhouser defines *avoidance behavior* as "avoiding evidence that not-p in a way that shows the agent already possesses sufficient information that not-p" (2005, p. 309). He offers the example of a subject who is self-deceived about her husband's affair and who is reluctant to drive past his purported lover's driveway for fear of seeing his car parked there. Tamar Gendler offers the example of a subject who is self-deceived about their child's innocence and who is reluctant to read newspaper reports for fear of coming across information linking their child to a crime (2007, p. 244). Rationalizers may engage in avoidance behaviors

through characteristic obfuscatory tactics designed to draw attention away from the flimsiness of their accounts. When their fragile accounts are challenged, they may change the subject or simply refuse to engage. Another strategy available to rationalizers is to screen their interlocutors, avoiding individuals who are critically minded or differently motivated in order to shield their sham reasoning from exposure. These kinds of behaviors suggest a responsiveness on the part of rationalizers to the flimsiness of their reasoning. This responsiveness is constitutive of the "suspicion" that is built into the non-naïveté constraint.[4]

A second source of evidence that rationalizers suspect that the considerations they adduce fail to establish the conclusions they reach comes from a family of cases that Gendler (2007) refers to as "trumped incentive." In these cases, some other goal comes to matter more to the subject than the goal of maintaining the impression of being "in a not-P world," and consequently they are willing to allow the rational belief P to play its "rightful thought-occupying and action-guiding role" (Gendler, 2007, p. 244). For example, suppose a medicine becomes available that will confer dramatic benefits to a disease sufferer who denies they are sick. In such a context of a high-stakes forced choice, the person will likely take the medicine, acting on the rationally based belief rather than on the imagining. Funkhouser (2005) makes a similar point: "the behavioral dispositions of the self-deceived, especially when in situations where the costs of mistake are high, are tipped toward believing the truth" (p. 307) We can apply the same analysis to cases of rationalization. In some (but not all) cases in which the stakes are raised and the valence of incentives is reversed, the rationalizer will abandon their rationalizing postures (although they may well slip back into their former ways if the situation changes again).

The strategic avoidance of rationalizers together with their changeability when incentives are reversed makes sense of our contemptuous rather than exculpatory attitude toward them. We intuitively recognize this kind of rationalization as something that rationalizers do, not merely an infelicitous influence on their belief-forming mechanisms that befalls them. Granted, my regimentation of the term *rationalization* to refer to the set of cases where the rationalizer is non-naïve is just a terminological stipulation. But I think

[4] None of this is to deny that in many cases the ultimate result is that the subject ends up believing the conclusion they have reached via rationalization. At this stage (which, following Funkhouser, we might label *self-delusion*), we should not expect the same strategic avoidance of evidence that we find in rationalizers.

that these cases of pretend inquiry are particularly conceptually and psychologically interesting and that they represent a neglected feature of our mental life that is worth paying attention to. The central case of rationalization I describe in this chapter—that of Jane Austen's John and Fanny Dashwood from *Sense and Sensibility*—is a vivid paradigm of this interesting subspecies of rationalization. This episode of pretend inquiry enables the Dashwoods to feel self-righteous about behavior that they themselves suspect is disgraceful.

The third thesis, deliberative weighing, is illustrated by contrasting rationalizers with subjects who avow belief for reasons that are manifestly arbitrary. Adam Elga (2005) offers the example of his friend, Daria, who believes in astrology and clings to her belief in defiance of the evidence. When Daria is confronted with evidence that her belief in astrology was unfounded, she concedes that she is unable to find any contrary evidence to support her belief in astrology. The sole reason she provides for sticking to her guns is "Believing in astrology makes me happy" (p. 115). Daria's endorsement of the truth of astrology is totally unresponsive to her assessment of the evidence.[5] As a result, she feels no pressure to provide reasons for endorsement. This distinguishes her case from that of the rationalizer, who does feel the pressure to appear responsive to evidence and whose rationalization is constructed so as to cultivate this appearance.

My proposal for reconciling the three constraints is to model the rationalizer as engaging in imaginative pretense. Rationalizers make as if they are guided by the aim of truth (believe *p* only if *p*), when in fact they are guided by a related but distinct aim, the aim of plausibility. This latter aim requires only that the considerations that rationalizers adduce in support of their conclusions have the appearance—to the rationalizers themselves and sometimes to others—of constituting sufficient reason. Deliberative exclusivity is respected because rationalizers are modeled not as believers but rather as pretenders. Guidance by the aim of plausibility is compatible with the suspicion that the considerations do not in fact establish the relevant conclusion. Rationalizers are not naïve. Deliberative weighing is satisfied because the product of rationalization bears an essential relation to the consideration that the rationalizer adduces. Rationalizers do not believe at will; rather, they engage in a pretense that is constrained by the evidence only indirectly.

[5] Some readers will be skeptical that Daria really believes in the validity of astrology. I do not weigh in on this question. See Huddleston (2012).

Following Liao and Gendler (2011), I use the term *imagination* to refer to our capacity to simulate different perspectives. This sense of (*recreative*) *imagination* is distinct from what Currie and Ravenscroft (2002) call *sensory imagination*[6] (the willful capacity to have perception-like experience in the absence of relevant stimuli) and *creative imagination*[7] (the capacity to combine ideas in unexpected and unconventional ways). *Imaginative pretense* or *pretense* refers to the guidance of action (including mental actions, such as mock inquiry) by imagination in the first sense. The strength of an individual's creative imagination is associated with the range of things they are able to imagine in the first sense, and hence with their degree of adeptness as a rationalizer.

The opening pages of Jane Austen's *Sense and Sensibility* contain a paradigm of the kind of rationalization I am interested in.[8] Here is the context: When Mr. Dashwood dies, his estate passes directly to his only son, John Dashwood. Mr. Dashwood's second wife and their daughters, Elinor, Marianne, and Margaret, are left only a small income. On his deathbed, Mr. Dashwood elicits a promise from his son, John, to use his inherited fortune to take care of his half- sisters. John and his wife Fanny consider the matter of exactly how much is owed to the half-sisters. Earlier, John had decided that a lump sum of £3,000, as recommended by his father, would ensure his sisters' financial security and discharge his promissory obligation to his father. Over the course of a conversation with his wife, however, his just and magnanimous feelings give way to his wife's meanness. After protracted rationalization, he decides that no more is owed to the sisters than "neighborly acts." By the end of their conversation, Fanny and John are collusive co-rationalizers.

The diminishment of the requital from £3, 000 to mere "neighborly acts" is accomplished through masterful rationalization. Austen's dialogue illustrates the distinctive repertoire of strategies that rationalizers deploy to reach their desired conclusions. Rationalizers tend to adduce pseudo-reasons, considerations that have only the appearance of relevance to the deliberative question. For instance, Fanny objects to paying the sisters an annuity because "it raises not gratitude at all." (Of course, the question of what would make the sisters feel grateful is orthogonal to the question of what is owed to them.) John Dashwood chimes in with a pseudo-reason of his own, insisting that an

[6] Van Leeuwen (2013) calls something similar "imagistic imagining."
[7] Van Leeuwen (2013) calls something similar "constructive imagining."
[8] Quotations are drawn from the Project Gutenberg reproduction found here: http://www.gutenberg.org/ebooks/161.

annuity "would only enlarge their style of living." (Here is another irrelevant consideration.) The Dashwoods also adduce weak reasons, considerations that are relevant to the question at hand but that are given undue weight. For instance, Fanny asks how, in eliciting such a promise, John's father "could answer it to himself to rob his child, his only child too, of so large a sum?" Fanny makes the hilarious conjecture that "people always live forever when there is an annuity to be paid them." A common strategy of rationalizers is to support their conclusions with empirical claims that are difficult to verify or to falsify.

What all of these strategies have in common is that they foster the appearance of sound and ineluctable reasoning while still affording crucial flexibility regarding the conclusion. I propose that rationalization should thus be modeled as the negotiation of two compatible but interacting aims: the aim of reaching a conclusion that is desirable and the aim of getting there with a story that is believable. These aims, taken individually, are in some instances pursued suboptimally. It may be the case that the rationalizer is unable to construct a sufficiently plausible account that leads to the most desirable conclusion. Rationalizers who are adept at the cultivation of suspension of disbelief will be able to arrive at conclusions that are relatively far-fetched. John and Fanny Dashwood are maestros in this regard.

Rationalization, Imaginative Engagement, and Emotion

For an episode of rationalization to mitigate feelings of guilt or cultivate feelings of self-righteousness, the content of the rationalization must be believable. But by *believable* I do not mean that the rationalizer is "able to believe" the story they tell. Fanny Dashwood, for example, is far too intelligent and far too shrewd to be taken in by an account that is so flimsy. We can be confident that if for some reason the conclusion Dashwood reaches were no longer valuable to her, she would quickly jettison the story. Dashwood's story is believable in sense in which we say that a novel or a film's plot is believable; that is, it is amenable to being richly imagined, to being imagined in a way that engages emotion and desire. As the rationalization unfolds, Dashwood is able to cultivate felt responses of self-righteousness and even of indignation.

Although most theorists think that there is a cognitive component to emotion, this does not imply that emotions require beliefs about their objects. The cognitive component of emotion can consist of a variety of different kinds of thoughts, which may include beliefs as well as construals, seeings-as,

entertainings, and imaginings. Noël Carroll describes well the way in which merely entertaining a certain kind of thought is often sufficient to generate the experience of fear or squeamishness in a subject.

> While cutting vegetables, imagine putting the very sharp knife in your hand into your eye. One suddenly feels a shudder. You need not believe that you are going to put the knife into your eye. Indeed, you are not going to do this. Yet merely entertaining the thought, or the propositional content of the thought (that I am putting this knife into my eye), can be sufficient for playing a role in causing a tremor of terror. (2001, p. 234)

We experience a rich variety of emotions when we think about past, future, possible, or even impossible states of affairs. For example, we may experience embarrassment in the absence of any belief that we have done something humiliating. Just as the phenomenon of the "near miss" can occasion fear, so also can it spur mortification. It is a common experience to think back on something that you came close to saying or doing and to feel acute embarrassment. We respond emotionally to both the actual and the possible faux pas (although not necessarily with the same intensity). Once we pay attention to the fact that thoughts directed toward mere *possibilia* form a significant component of our emotional lives, it should seem less mysterious that the pretense involving activity of rationalization has the emotional payoff that it does.

Work on the neuroscience of emotion suggests that imaginary stimuli lead to emotional effects via the same kind of causal pathways that generate emotion from real stimuli (Schroeder & Matheson, 2006). This picture helps us to understand why subjects exhibit such strong emotions toward narratives that they know are not real. Schroeder and Matheson (2006) trace a causal network between tokenings of an imagined representation, on the one hand, and strong feelings, on the other, concluding that there is all but decisive evidence in favor of the view that the exercise of imagination is capable of eliciting real and vivid emotion. This supports the view that imaginative acts have the power to move us emotionally through the activation of what they term a "distinct cognitive attitude."

The kind of imaginative engagement we experience when immersed in a vivid and arresting narrative engages desire and emotion.[9] While some have expressed skepticism about whether such emotions are genuine and rational,

[9] Gregory Currie distinguishes "suppositional imagining" from "rich imagination" that has a "desire-like component" (2002, 215).

no one denies that narratives can elicit strong feelings even when we are quite aware that the narratives are fictional.[10] I propose that there is an illuminating parallel to be drawn between the emotions experienced by readers of fiction and film spectators in response to narrative and the pretense-based emotions implicated in rationalization. Once we notice that merely entertaining a thought can elicit a strongly felt emotional response, it should seem less strange that rationalizers, who need not believe the content of the rationalizations, may nevertheless be moved by them emotionally.

Another relevant feature of emotion is the tendency of emotions to "spill over" from objects for which the emotion is fitting to objects that merely resemble objects for which the emotion is fitting. Patricia Greenspan (1988) describes a subject who, having been bitten by a rabid dog in the past, is now deathly afraid of Fido, a harmless old hound that is well known to him. Although he feels fear in the presence of Fido, he knows very well that Fido will not hurt him. This belief is in evidence when he doesn't shield his children from Fido or run away screaming. What is interesting about the example is that fear described does have a cognitive element; it is not purely physiological. Yet, it seems wrong to impute to the subject a belief that he is endangered by Fido:

> Instead of supposing that his beliefs come into momentary conflict whenever Fido comes near, it seems simpler, and preferable from the standpoint of rational explanation, to take this as a case where emotion parts from judgment. It exhibits the tendency of emotions, in contrast to a rational agent's beliefs, to spill over to and to fix on objects resembling their appropriate objects in incidental ways. (Greenspan 1988, p. 18)

The "spill over" effect that Greenspan identifies may shed more light on the imaginative character of rationalization. Rationalization allows us to feel emotions associated with the belief that there are exculpating circumstances in response to considerations that merely appear to be exculpatory.

Similarly, we may feel indignation in response to considerations that merely resemble what we take to be good grounds for indignation (recall Fanny Dashwood's emphatic insistence that paying an annuity will raise no gratitude). The phoniness of these emotions does not threaten their felt

[10] See Radford (1975) for a defense of the view that fiction-directed emotions are irrational; see Walton (1978) for a defense of the view that fiction-directed emotions are not genuine emotions.

intensity. Rationalizations can vary in how successfully they mitigate feelings of guilt and engender feelings of self-righteousness. Well-constructed rationalizations succeed in silencing the voice of conscience. Rationalizations that are wildly implausible or include specific and falsifiable claims are vulnerable to debunking considerations that work to undermine the suspension of disbelief.

Here one might object that it is simply implausible to suppose rationalizers are engaging in imaginative pretense since the experience of rationalizing certainly feels rather different from acting in a play or participating in a game of make-believe. But just as a method actor who is preparing for a role can pretend that certain things are true of their life without consciously attending to the fact that this is what they are up to, so too the rationalizer may pretend that certain considerations provide conclusive reasons for belief. The key difference between the rationalizer and the method actor is that the method actor has at an earlier time consciously framed their activity as one of pretense, whereas the rationalizer has not.

In paradigm instances of imaginative pretense, the subject's beliefs and desires play a variety of roles in shaping pretend behavior. Van Leeuwen (2011) describes the manner in which cognitive and conative attitudes "comment on and constrain imaginings," thereby influencing pretend action (p. 67). He notes that, in particular, they can comment on "the *value* to the agent of a particular imagined action" (p. 67, italics in original). We might add that they can further comment on the value of pretending in one particular way rather than another and on the value of initiating or terminating pretense at one time rather than another. Even pretenders who are deeply immersed in an imaginative project (method actors, for instance) keep track of their mental attitudes in this way. Liao and Doggett (2014) cite supporting evidence from developmental psychology indicating that children as young as 3 years keep track of the fact that they are pretending while engaged in immersive fantasy play even in the face of adult intervention designed to blur the boundary between fantasy and reality.[11] For instance, when an adult actually bites into a Playdough cookie (as opposed to merely pretending to bite), children are clearly shocked by the transgression ("Oh, you took a real bite. Now your teeth are all pink. How does it taste? . . . Yuck, do you always eat Playdough?") Here,

[11] The original study was carried out by Golomb and Kuersten (1996). Liao and Doggett cite a description of the study in Taylor (1999).

the child's shock indicates that despite their immersion in the imaginary game, they never lose track of the fact that they are only pretending that the "cookies" are edible.

In contrast to these witting and self-aware pretenders, my hypothesis is that the rationalizer has diminished access to the metacognitive mental content (S pretends that *p*). Although the representation that they pretend that *P* rather than believe that *p* may in some contexts guide their action (e.g., when they engage in characteristic evasive maneuvers), it is not readily accessible. The value of the pretense, its appropriate manner, and its proper limits are therefore not the object of comment and constraint. The fact that the rationalizer does not direct rational scrutiny to the circumscription of contexts in which they are guided by imagination rather than by belief explains why the behavioral and emotional responses to the pretense are not quarantined as they are in the imaginative pretense of actors and role-players. A rationalizer may stick with a rationalization even when so doing becomes self-undermining.

Rationalization and Imaginative Resistance

Tamar Gendler (2000) coins the term *imaginative resistance* to refer to the blockages that subjects sometimes encounter in their engagement with prompted imaginative activities. For example, film audiences and readers of fiction may "pop out" of fictional worlds in which deviant moral codes appear to operate. Moran (1994) offers an example in which you are confronted with a variation of *Macbeth* where "the facts of [Duncan's] murder remain as they are in fact presented in the play, but it is prescribed in this alternate fiction that this was unfortunate only for having interfered with Macbeth's sleep." In a similar vein, Kendal Walton's (1994) one-line story ("In killing her baby, Giselda did the right thing; after all, it was a girl") is a starker example of the barriers to imaginative engagement. Readers and spectators of these morally distorting fictions will "pop out" and fail to feel the emotions mandated by the narrative.

In these highly schematic fictional narratives, the moral incoherence of the story is blatant. But an author who is skilled enough to disguise the moral incoherence of a narrative can bring us to respond emotionally to fictional states of affairs that would otherwise present resistance. In much the same way, a maestro rationalizer's tale of exculpation can be far-fetched in

proportion to the measure in which it is carefully crafted to paper over moral and logical incoherence.

Walton (1994) famously asked, "There is science fiction, why not morality fiction?" (p. 37). We should reject the question's presupposition. Genres like the revenge epic and the gangster film are a kind of morality fiction insofar as they allow us to respond emotionally in ways that are inconsistent with our considered moral judgments. For a short time we are able to inhabit a world where retribution need not be mitigated by mercy or where fealty to omertà dominates all other values. In our imaginative engagement with revenge epics and gangster films, the reader or viewer is often complicit in bringing about a divergence between occurrent moral emotion and settled moral judgment. It is within the power of a sufficiently skilled author to create the conditions of this kind of "imaginative promiscuity." In this vein, Meskin and Weinberg observe that "it is a noticeable feature of artistic practice that talented authors can turn the unimaginable into the stuff of fiction" (2011, p. 240).

Just as moral incoherence may be disguised, so too rational incoherence may be disguised, thereby dislodging further blockages to rich imaginative engagement. Gendler (2000) presents a story, "The Tower of Goldbach," that contains the conceptual impossibility "twelve both is and is not the sum of five and seven" (p. 66). Despite the conceptual incoherence at its core, the story is able to nonetheless sustain our imaginative engagement.[12] Gendler's strategy is to get the reader to focus on certain elements of the story and thereby ignore others.[13] In particular, the reader is led away from attending to the conceptual impossibility, which is deftly hidden in the story: Gendler relies on obfuscation to make the story work. As Meskin and Weinberg (2011) put it, "what was once impossible-and-unimaginable is rendered impossible-yet-somehow-imaginable" (p. 248).

Meskin and Weinberg (2011) point out that Gendler's story demonstrates the "nonmonotonicity" of imaginative blockage: "One can have a fiction with blockage, and add more fictional contents to it *even without rendering the imagined contents consistent*, and not necessarily end up with blockage in the new fiction" (p. 247, italics in original). Similarly, the convoluted elaborations of rationalizers do not render their stories morally or rationally coherent. Rather, they serve to disguise incoherence and thereby secure emotional

[12] Graham Priest (1997) does something similar in "Sylvan's Box," a story about an absolutely empty box that has something in it.

[13] Stock (2003) expresses skepticism regarding whether Gendler really gets the reader to imagine a conceptual impossibility.

uptake. Rationalizers may face the same kind of imaginative blockages experienced by audiences of standard narratives. Journeymen rationalizers will be stymied by such imaginative obstructions; maestro rationalizers have the skills to work around them.

In the example from Austen, John Dashwood experiences a kind of imaginative blockage. Hesitant in the face of his wife's unwholesome purposes, John Dashwood insists that he will not break his promise to his father: "The promise, therefore, was given, and must be performed." Fanny Dashwood manages to draw John's attention away from the morally troubling breach of promise and toward consideration of the welfare of his son, allowing John to feel righteous when he ought to be ashamed:

> Well, then, let something be done for them; but that something need not be three thousand pounds. Consider . . . that when the money is once parted with, it never can return. Your sisters will marry, and it will be gone forever. If, indeed, it could be restored to our poor little boy.

There is also something rather impressive (though doubtless morally pernicious) about what Fanny does here. Austen gives us the distinct impression that the ingenuity of Fanny's story manifests a kind of cognitive achievement.

This virtuosic aspect of the rationalizing mind is featured prominently in the work of Dan Ariely on the rationalization of dishonesty. Ariely proposes that "the link between creativity and dishonesty seems related to the ability to tell ourselves stories about how we are doing the right thing, even when we are not" (2012, p. 197). He proposes that "the more creative we are, the more able to come up with good stories that help justify our selfish interests" (p. 197).

The pretense account of rationalization allows us to understand better the link that Ariely finds in the experimental data. We should not think of pretense as offline processing that is altogether segregated from belief and devoid of motivational force. Nichols and Stich (2000), for example, explicitly set out to provide a model of pretense that explains how it is that "the events that occurred in the context of pretense have only quite limited effect on the post-pretense cognitive state of the pretender" (p. 120). But this kind of "quarantining" of pretense is conspicuously absent in rationalization. As Gendler (2003) points out, in certain contexts quarantining gives way to its opposite—"contagion"—whereby the pretended contents come to be believed, or treated as if they are believed, merely because they are pretended.

In cases of "affective transmission," mere contemplation of a content that is emotionally charged causes the thinker to behave and feel in a way that is consistent with belief in that content (p. 131). So, for example, we can imagine that Mr. Dashwood may respond with genuinely felt umbrage and affront if he is later accused of reneging on his promise.

The model of rationalization as pretense also tells against models of pretense whereby the pretender processes belief-eligible content in the same way that they process belief. For example, the Nichols and Stitch model of pretense sets out to explain how it is that "inference mechanisms treat pretense representations in roughly the same way that the mechanisms treat real beliefs" (2000, p. 125). Although Nichols and Stitch are right that such "mirroring" is typical, Gendler points out that pretense episodes may also manifest "disparity," the tendency whereby pretense content differs from non-defective belief content in that what is pretended may be incomplete (some features may remain permanently unspecified and unspecifiable) as well as incoherent (some features may be logically and conceptually incompatible) (2003, p. 137). The contents of rationalization are typically incomplete (they do not establish the conclusion that is reached), and the considerations adduced are often irrelevant. But neither of these features is an insuperable impediment to imaginative and emotional engagement.

Conclusion

The species of rationalization that is my explanandum bears a striking resemblance to what Harry Frankfurt (2006) famously dubbed "bullshit." Frankfurt gives a characterization of bullshit as a contrast to lies. When one tells a lie, one deliberately tries to cause another person to believe something that one takes to be false. When one merely produces bullshit, one misleads another person as to what one is up to. To illustrate the distinction, Frankfurt offers the example of a "Fourth of July Orator" who waxes bombastic about "our great and blessed country, whose Founding Fathers under divine guidance created a new beginning for mankind" (pp. 120–121). The orator is not lying since he is not concerned with bringing about false beliefs in his audience about the role of the deity in founding the country; he is uninterested in his audience's historical or theological views. Rather, the orator is trying to convey a certain impression of himself as a patriotic man. The orator merely makes as if he is trying to convey information about the founding fathers. For

Frankfurt, bullshitting "unfits" a person for the truth by fostering a habitual indifference to it.

We can think of rationalizers on the model of "self-bullshitters": They are Frankfurtian bullshitters who bullshit others as well as themselves. Rationalization and bullshit both involve the use of misdirection through pretending. Bullshitters make as if they are concerned with conveying the content of what they say, when in fact they are merely trying to convey a certain impression of themselves. Rationalizers make as if their aim is honest inquiry, when in fact it is only plausibility and self-justification. Like expert rationalizers, expert bullshitters exercise skill in crafting their bullshit so that it is not easily detected and not easily debunked. Both figures can be understood as engaging in a kind of pretense (although only the bullshitter requires an audience apart from themselves). The bullshitter succeeds if they manage to convince an audience that their narrative is in fact expressive of the person they really are. The rationalizer succeeds if their narrative creates for them an imaginative experience that is rich enough to make them feel better about themselves or that soothes their conscience or that allows them to act in ways they know or suspect to be wrong.

Morally distorting narratives, if crafted with care, can bring us to root for the bad guy. Rationalization can be understood as a kind of morally distorting narrative where the "bad guy" is the rationalizer themselves. The intended audience for the story is sometimes other people but most often also the rationalizer themselves. The rationalizer spins a tale for themselves whose major theme is their own exculpation and self-justification.

Although our emotional response to fiction can sometimes tell of our real commitments, this is not always the case. With genres like the revenge epic, we allow ourselves some latitude in letting our emotional response part ways from our settled judgment because we know that there is less at stake. (Fictional characters cannot be harmed.) As a result, there is a sense in which fictional narratives allow us a "safe space" to experience the kind of emotional responses that issue from a moral point of view that our reflective selves firmly repudiate. Engagement in fictional worlds makes possible a kind of emotional promiscuity.

Rationalization, too, opens up a space for emotional promiscuity. But the rationalizing stories we tell have implications in the real world, sometimes very serious implications. Although I have characterized rationalization as a sometimes virtuosic cognitive and creative achievement, I am in no measure condoning the morally dubious behavior that rationalization brings in its

wake.[14] In Austen's fiction, rationalization is what catalyzes John Dashwood's unjustifiable breach of promise. Social scientists find similar patterns. For instance, Wegner et al. (2015) find that increased use of justification by perpetrators of sexual aggression is a significant predictor of further sexual aggression. The role that rationalization plays in the deformation of practical rationality and of moral character is a rich terrain for future inquiry. Essential to being a cognizer who can rationalize is a capacity for the kind of vivid imagination that engages emotion and desire.

Acknowledgments

This chapter has benefited from written comments by Nathan Ballantyne and David Dunning, as well as oral feedback from the participants in the NYC Epistemology and Psychology Conference, June 2016.

References

Ariely, D. (2012). *The honest truth about dishonesty: How we lie to everyone—Especially ourselves*. HarperCollins.

Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 4, pp. 1–34). Academic Press.

Boghossian, P. (2003). The normativity of content. *Philosophical Issues*, *13*(1), 31–45.

Carroll, N. (2001). *Beyond aesthetics: Philosophical essays*. Cambridge University Press.

Currie, G. (2002). Desire in imagination. In T. Gendler & J. Hawthorne (Eds.), *Conceivability and possibility* (pp. 201–222). Oxford University Press.

Currie, G., & Ravenscroft, I. (2002). *Recreative minds: Imagination in philosophy and psychology*. Oxford University Press.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, *60*(23), 685–700.

D'Cruz, J. (2014). Rationalization as performative pretense. *Philosophical Psychology*, *28*(7), 980–1000.

D'Cruz, J. (2015). Rationalization, evidence, and pretense. *Ratio*, *28*(3), 318–331.

Elga, A. (2005). On overrating oneself . . . and knowing it. *Philosophical Studies*, *123*(1–2), 115–124.

Engel, P. (2013). Doxastic correctness. *Aristotelian Society Supplementary Volume*, *87*(1), 199–216.

Foley, R. (1993). *Working without a net: A study of egocentric epistemology*. Oxford University Press.

[14] For an elaboration of the potential moral benefits of rationalization, see Summers (2017).

Frankfurt, H. (2006). *On bullshit*. Princeton University Press.
Funkhouser, E. (2005). Do the self-deceived get what they want? *Pacific Philosophical Quarterly*, *86*(3), 295–312.
Gendler, T. (2000). The puzzle of imaginative resistance. *The Journal of Philosophy*, *97*(2), 55–81.
Gendler, T. (2003). On the relation between pretense and belief. In D. McIver Lopes & M. Kieran (Eds.), *Imagination, philosophy, and the arts* (pp. 125–141). Routledge.
Gendler, T. (2007). Self-deception as pretense. *Philosophical Perspectives*, *21*(1), 231–258.
Gendler, T. S., & Liao, S.-y. (2016). The problem of imaginative resistance. In J. Gibson & N. Carroll (Eds.), *The Routledge companion to philosophy of literature* (pp. 405–418). Routledge.
Gino, F., & Ariely, D. (2012). The dark side of creativity: Original thinkers can be more dishonest. *Journal of Personality and Social Psychology*, *102*(3), 445–459.
Glüer, K., & Wikforss, A. (2009). The truth norm and guidance: A reply to Steglich-Petersen. *Mind*, *119*(475), 757–761.
Golomb, C., & Kuersten, R. (1996). On the transition from pretence play to reality: What are the rules of the game? *British Journal of Developmental Psychology*, *14*(2), 203–217.
Greenspan, P. (1988). *Emotions and reasons: An inquiry into emotional justification*. Routledge, Chapman and Hall.
Hazlett, A. (2013). *A luxury of the understanding: On the value of true belief*. Oxford University Press.
Holland, N. (2008). Spider-Man? Sure! The neuroscience of suspending disbelief. *Interdisciplinary Science Reviews*, *33*(4), 312–320.
Huddleston, A. (2012). Naughty beliefs. *Philosophical Studies*, *160*(2), 209–222.
Liao, S.-y., & Doggett, T. (2014). The imagination box. *The Journal of Philosophy*, *111*(5), 259–275.
Liao, S.-y., & Gendler, T. (2011). Pretense and imagination. *Wiley Interdisciplinary Reviews*, *2*(1), 79–94.
McHugh, C. (2012). The truth norm of belief. *Pacific Philosophical Quarterly*, *93*(1), 8–30.
Meskin, A., & Weinburg, J. (2011). Imagination unblocked. In E. Schellekens & P. Goldie (Eds.), *The aesthetic mind: Philosophy and psychology* (pp. 239–253). Oxford University Press.
Moran, R. (1994). The expression of feeling in imagination. *The Philosophical Review*, *103*(1), 75–106.
Nichols, S., & Stich, S. (2000). A cognitive theory of pretense. *Cognition*, *74*(2), 115–147.
Owens, D. (2000). *Reason without freedom: The problem of epistemic normativity*. Routledge.
Owens, D. (2003). Does belief have an aim? *Philosophical Studies*, *115*(3), 283–305.
Priest, G. (1997). Sylvan's box: A short story and ten morals. *Notre Dame Journal of Formal Logic*, *38*(4), 573–582.
Radford, C. (1975). How can we be moved by the fate of Anna Karenina? *Proceedings of the Aristotelian Society Supplementary Volume*, *49*, 67–93.
Schroeder, T., & Matheson, C. (2006). Imagination and emotion. In S. Nichols (Ed.), *The architecture of the imagination: New essays on pretence, possibility, and fiction* (pp. 19–40). Oxford University Press.
Shah, N. (2003). How truth governs belief. *The Philosophical Review*, *112*(4), 447–482.
Shah, N., & Velleman, J. D. (2005). Doxastic deliberation. *The Philosophical Review*, *114*(4), 497–534.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science*, *24*(2), 125–130.

Steglich-Petersen, A. (2009). Weighing the aim of belief. *Philosophical Studies*, *145*(3), 395–405.

Stock, K. (2003). The tower of Goldbach and other impossible tales. In D. McIver Lopes & M. Kieran (Eds.), *Imagination, philosophy, and the arts* (pp. 107–124). Routledge.

Summers, J. (2017). *Post-hoc ergo propter hoc*: Some benefits of rationalization. *Philosophical Explorations*, *20*(Suppl. 1), 21–36.

Taylor, M. (1999). *Imaginary companions and the children who create them*. Oxford University Press.

Van Leeuwen, N. (2011). Imagination is where the action is. *The Journal of Philosophy*, *108*(2), 55–77.

Van Leeuwen, N. (2013). The meanings of "imagine": Part 1: Constructive imagination. *Philosophy Compass*, *8*(3), 220–230.

Velleman, J. D. (2000). *The possibility of practical reason*. Oxford University Press.

Walton, K. (1978). Fearing fictions. *Journal of Philosophy*, *75*(1), 5–27

Walton, K. (1994). Morals in fiction and fictional morality. *Proceedings of the Aristotelian Society Supplementary Volume*, *68*, 27–50.

Wedgwood, R. (2013). Doxastic correctness. *Aristotelian Society Supplementary Volume*, *87*, 217–234.

Wegner, R., Abbey, A., Pierce, J., Pegram, S. E., & Woerner, J. (2015). Sexual assault perpetrators' justification for their actions: Relations to rape supportive attitudes, incident characteristics, and future perpetration. *Violence Against Women*, *21*(8), 1018–1037.

6

Sight Unseen, Justice Unobserved

How Naïve Realism in Visual Attention Affects Legal Decision-Making

Yael Granot, Kristyn A. Jones, and Emily Balcetis

In August 2014, Michael Brown, an 18-year-old Black man from Ferguson, Missouri, was fatally shot by police officer Darren Wilson. Accounts of events leading up to the shooting varied wildly. As Brown could not speak for himself, investigators relied largely on witness statements, which portrayed a struggle in which Officer Wilson chased and shot Brown. In the courtroom, much of the weight of the evidence turned on the testimony of Officer Wilson. Wilson described grappling with Brown, who he said was reaching for Wilson's gun (Bosman et al., 2014). Based upon these narratives, the grand jury found Officer Wilson not guilty of any wrongdoing, a decision that incited several nights of protests in Ferguson as well as cities across the country ("Ferguson and Other Cities React," 2014). Public outcry in response to these events suggested that video evidence would have facilitated justice in the case. Reports asserted, "a camera on Wilson's uniform would have ended the uncertainty and potentially avoided the subsequent tumult that engulfed the St. Louis suburb" (Sanburn, 2014).

Yet, later that same summer, in Staten Island, another Black man, Eric Garner, was choked to death in an encounter with police. Cell-phone video footage captured by a witness to the incident documented the altercation in its entirety. To many Americans the video overwhelmingly proved police guilt; 68% of New Yorkers polled in the following month said that there was no excuse for the police actions against Garner (Carroll, 2014). However, the grand jury in the subsequent case against the police evaluated the same video evidence and ruled otherwise, failing to indict the officer. Public reactions to this decision largely took the form of shock and distress. People presumed that with the available video everyone would see the same truth of what happened with their own eyes (Capehart, 2015). Rather than unifying the

Yael Granot, Kristyn A. Jones, and Emily Balcetis, *Sight Unseen, Justice Unobserved* In: *Reason, Bias, and Inquiry.* Edited by: Nathan Ballantyne and David Dunning, Oxford University Press.
DOI: 10.1093/oso/9780197636916.003.0007

public's understanding of events, the available video documentation—the main evidence available to the public—actually polarized opinions about the court's decision among different social groups. Indeed, 90% of Black Americans surveyed but only 47% of White Americans contested the grand jury's ruling in the Garner case (Pew Research Center, 2014). Video evidence did not eliminate differences in how those who viewed it understood the case facts. Instead, despite watching the same evidence, people came to starkly different legal judgments.

The juxtaposition of these two events within the very same year serves as a powerful refutation of the idea that video evidence is a panacea to legal bias. People do not always see the same objective truth and video footage does not always bridge divides. In a more recent example, Ma'Khia Bryant, a 16-year-old Black girl who was wielding a knife, was fatally shot when officers arrived on scene (Alonso & Sutton, 2021). Some viewers who watched the 15-second body-worn camera video of the event focused on the fact that Bryant was holding a knife, leading them to hail the officer a hero (Mansfield & Kenton, 2021). Others, however, argued that the officer acted negligently putting bystanders' lives in danger when he fired his weapon, and that de-escalation would have been a more appropriate response (Cineas, 2021; Morris, 2021).

Yet certainly, video evidence can motivate action in response to bias; videos capturing the killing or assault of Black Americans like George Floyd, Laquan McDonald, and Sandra Bland at the hands of police, fomented calls for accountability and broader systemic change. This may be why national polling consistently finds overwhelming, universal support for body cameras, regardless of the respondent's race: 93% of White Americans and 93% of Black Americans favored requiring police officers to wear body cameras while on duty (CBS News/New York Times, 2015; for a discussion, see Sommers, 2016). Legal experts echo this lay faith in video evidence. A 2013 judicial ruling in response to the New York City Police Department's stop-and-frisk policy required the city to conduct a pilot program with body cameras, suggesting that they would "provide a contemporaneous, objective record" that "may help lay rest to disagreements that would otherwise remain unresolved" (*Floyd v. City of New York*, 2013). Embedded within this ruling and the public's support for cameras are two significant assumptions. First, people believe that video evidence offers a complete, objective, and unmediated record of events. Second, people believe that they are capable of seeing this evidence fully and objectively.

In this chapter, we call into question these assumptions. Drawing from decades of research across the social, cognitive, and vision sciences, we explore the limitations of video and the way in which individuals view it. These limitations and subjective viewing experiences together combine to challenge individuals' abilities to form a complete and accurate understanding of events depicted within. We confine our analysis to the impact of selective visual attention and discuss how people gather visual information in biased ways. We also describe people's failure to scrutinize their beliefs about what they think they see. As a result, they hold a deep-seated faith in the reliability of their visual experiences and the erroneous belief that what they see is always an accurate and complete representation of the world. These tenets of visual experience do more than undermine lay intuitions about how perception functions. When they manifest during legal proceedings, they may shake the foundations of justice. We conclude by exploring the ways in which biases in visual attention can powerfully undermine a legal system that is currently turning to visual evidence for help.

The Limitations of Video Evidence

Video evidence proliferates in legal arenas because of the belief that such material can provide a full, unmediated account of events. Rulings in high-profile legal cases support this conception. For example, in *Scott v. Harris* (2007), a police officer, Timothy Scott, rammed the back of civilian Victor Harris's car to end a high-speed chase, leaving Harris paraplegic. Harris claimed the officer's actions violated his Fourth Amendment rights against unreasonable seizure. Officer Scott claimed he was protecting the public from Harris's reckless driving. When the case reached the US Supreme Court, a majority of justices granted summary judgment in favor of the officer, based on their appraisal of dashboard camera video depicting Harris's driving. They wrote that no "reasonable juror" could dispute that Harris was driving recklessly enough to justify the officer's use of deadly force, citing the specific actions they saw in the video. Empirical work testing that assumption with the same video found stark group-based distinctions in whether people deemed there to be evidence of reckless driving (Kahan et al., 2009). Yet for the justices, the evidence of their own eyes, they wrote, was incontrovertible and universal.

The depiction of video in much of legal scholarship, argument, and rulings is one of a reliable, silent witness (Kaye et al., 2013). But many factors,

including the sheer constraints of video technology, belie that possibility. Video cannot capture everything in its entirety. The angle and distance of the camera from its target determine what information from a scene is captured and, just as importantly, what information is missing. For example, casinos are known to be some of the most highly surveilled public establishments. They have an "eye in the sky," probing for potential crime, danger, and disruption. Yet in October 2017, at the Mandalay Bay Resort and Casino in Las Vegas, Stephen Paddock shot a security guard and then perpetrated a mass shooting that would claim the lives of 58 people and leave hundreds more injured. Despite nearly 3000 cameras throughout the hotel, no cameras monitored the hallway where Paddock first shot the guard on duty (Pearce et al., 2017). Beyond this single event, police body-worn cameras limit access to pertinent information because of their outward-facing view. Such positioning fails to capture the officer's movements or demeanor and may similarly miss peripheral information that was available to the officer during an encounter with a civilian (Stoughton, 2018). In these and other ways, cameras may cement into record only a fraction of the critical information available within a scene. This incompleteness becomes particularly problematic in the perceptual experience of the perceiver who may, for example, incorrectly fill in information that was missing in a way that fits with their expectations (Foley et al., 2007).

Cameras also present a temporally limited depiction of events, which may influence the judgments perceivers make. For example, officers equipped with body-worn cameras may be able to turn the cameras they wear on and off, cropping the temporal sequence at their personal discretion. Local journalists in Minneapolis, Minnesota, found that despite regulations mandating that officers turn on their cameras during all traffic stops and encounters involving criminal activity, in 1 month in 2017, officers recorded less than 20 minutes of footage per 8-hour shift, which many have argued is a much lower amount of content than should be expected (Editorial Board, 2017).

Such piecemeal documentation of police encounters can be particularly problematic in the context of fatal shootings. For example, in Chicago in 2016, 18-year-old Paul O'Neal was driving a reportedly stolen car, and officers engaged in pursuit; as O'Neal tried to run away on foot after a crash, he was fatally shot in the back by police. Critically, body-worn cameras captured the beginning of the chase but cut out before the final shot (Phippen, 2016). Similarly, in 1989, when New York City police interrogated five

African American teenagers about the assault of a jogger in Central Park for between 14 and 30 hours each, the video presented in court captured only the final moments in which confessions were elicited (Kassin, 2002). Compared to the information that people directly confront, they may fail to perceive or underweight evidence that is absent from the footage—even when that information is critical to a full understanding of events.

Further, decisions made by the operator of recording technology determine what information reaches the perceiver. As a classic example, in 1942 Benito Mussolini circulated an image of himself on horseback wielding a sword, in which the handler holding the horse steady for him was removed from the image. Italian citizens did not doubt the integrity of the image and so were dutifully awed by their leader (Farid, 2009). As noted by Errol Morris, acclaimed documentary filmmaker, "the whole act of creating a photograph is an act of cropping reality" (Morris, 2011, p. 165). The angle a director chooses similarly skews the content that perceivers receive. Indeed, work on the camera perspective bias has shown that police interrogation videos that focus exclusively on the suspect foster more certain judgments of the suspect's guilt relative to interrogation videos that depict only the officer or both actors (Ware et al., 2008). Similarly, research shows that compared to dashcam and surveillance footage, body-worn camera footage, where the officer is less visually salient, leads to lowered judgments of officer intent and more lenient punishment decisions (Turner et al., 2019; Jones et al., 2020).

Some newly emerging policy initiatives reflect a nascent understanding that video recordings can be critically incomplete in their representation of events. For example, legal scholars have begun to advocate long-form videotaping of interrogations as well as the automation of police body-worn camera activation, in order to combat fragmented documentation (Barr, 2017; Kassin et al., 2010). Yet, people may still not realize what is outside the frame or what happened before or after the available recording and may make decisions based on mistaken assumptions about what is not captured. In its incomplete depiction of events and the way that perceivers experience that missing information, video may preclude objective decision-making.

Error in Visual Experience

Beyond the limitations of video recording technology—and how it is used by the camera people operating it—perceivers themselves fail to encode their

visual experiences in full. Ample scientific evidence documents individual differences in perception that suggest bias in people's visual experiences.

But a note first on what constitutes bias, in our argument. Following a rich tradition of interpersonal perception research in social psychology (e.g., Kenny & Acitelli, 2001; Kruglanski, 1989; Kunda, 1990), we suggest that visual perception can be decomposed into and precisely defined by two related but non-redundant concepts of accuracy and bias.

Perceptual *accuracy* is defined as the ability to correctly identify the visual experience. For instance, within research testing perception of others' emotional expressions, accuracy is calculated by comparing the discrepancy between participants' judgments of emotion and a correct judgment of the target (Brady & Balcetis, 2015). "Correct" identifications can be gauged by comparing perceivers' identifications to the discrete emotion as determined by the researcher (e.g., Isaacowitz et al., 2007, for a review), as intended by a target (e.g., Salovey & Mayer, 1990), or as indicated by a targets' own assessment of their facial expressions when viewing themselves afterward as if a third-party observer (e.g., Zaki et al., 2009). Accuracy, in this way, reflects the discrepancy between participants' beliefs about what they saw against an objective truth, as established through varied means.

In comparison to accuracy, *bias* can be defined as a systematic tendency for the perceptual system to privilege certain classes of information over others. Within the context of emotion perception, again, bias is typically measured by comparing the frequency, speed, or intensity with which perceivers categorize emotional displays as positive rather than negative. For instance, bias can be measured as discrepancies in different participants' reaction times when identifying happy from sad expressions (Niedenthal et al., 2002). Bias can also be measured as discrepancies in the intensity perceivers ascribe to positive rather than negative emotion expressions (Joorman & Gotlib, 2006). Finally, bias can be measured as a discrepancy in the number of negative versus positive emotions recognized in mixed emotion stimuli (Bouhuys et al., 1999). Bias then quantifies differences in how readily fractions of the complete visual scene are consciously experienced and is described by the tendency of different perceivers to privilege certain visual information at the expense of other information that the stimulus actually contains.

One leading cause for perceivers' biased visual experience is systematic differences in visual attention. Human eyes point forward, which narrows the field of view compared to many animals, like mice, whose eye placement enables them to incorporate lateral visual information (Lukáts et al., 2005).

Coupled with the restricted scope of the visual field, attentional resources are finite. Visual acuity is highest in the fovea, which is limited to the central 2 degrees of the visual field (Rayner, 1998). All other visual input is encoded peripherally and therefore has much less precise detail. Because visual resources are taxed by the complexity of the surrounding environment, attention must be selectively oriented to some information at the expense of others (Treisman, 2006). As a result, where we direct our gaze shapes how acutely we see something and whether we see it at all.

Measures of eye-tracking uncovered systematic differences in attentional patterns across individuals. For instance, individuals who suffer from some form of eating disorder, such as anorexia or bulimia, were more likely than healthy individuals to attend to areas of their own bodies that they personally considered unattractive (Bauer et al., 2017; Roefs et al., 2008). People suffering from schizophrenia, a disorder that affects emotional processing, oriented their visual gaze toward emotional faces at the expense of attending to neutral faces, but were less likely than healthy controls to spend time attending to angry and sad faces (Jang et al., 2016). Individuals in a positive mood attended to positively valenced images more than negatively valenced images, an attentional preference not present for individuals not experiencing a positive mood (Wadlinger & Isaacowitz, 2006). Conversely, individuals with major depressive disorder selectively attended to sad faces, while healthy individuals did not (Gotlib et al., 2004). Psychological states of the perceiver shift the direction of attention and produce a selective and incomplete representation of the external world.

Moreover, attentional patterns reliably differ as a function of group membership. For example, individuals whose political ideology was more left-wing were more likely to direct eye gaze longer to political posters from a liberal party than from a conservative party (Marquart et al., 2016). Racial group membership also predicts discrepant patterns of attention (Sternisko et al., 2017). Participants watched footage of a physical altercation between a White police officer and a White civilian on the side of the road. In the struggle, the civilian bit the officer's hand, and the officer returned with a blow to the civilian's head. Eye-tracking results revealed that White participants looked significantly less often to the White police officer than did Asian participants.

Though the psychological mechanisms that gave rise to discrepant patterns of visual attention across groups were not tested, one potential mechanism explaining these discrepancies is people's expectation that out-groups pose

threats (Stephan & Stephan, 2000). Expectations of threat facilitate selective attention to the source of the threat. Indeed, threatening stimuli capture faster and sustain attention longer than non-threatening stimuli (e.g., Öhman et al., 2001). And those who are most likely to expect a threat are most likely to attend to it. For example, individuals who have high levels of trait anxiety are more likely than those with low anxiety to attend to cues of a threat (Goodwin et al., 2017). In a similar manner, the expectation of threats originating from intergroup contexts directs visual attention; people attend more quickly and for longer to out-group faces that are threatening. Black faces capture the attention of White participants faster and hold it for longer than White faces, especially among participants who more strongly associate *Black* with *threat* (Donders et al., 2008). Even trained police officers show selective attention to Black rather than White faces; when tested within their own precincts, officers who were primed with words related to criminality initially directed attention toward the face of a Black civilian rather than a White one. This attentional bias did not emerge among officers for whom thoughts about criminality were not activated (Eberhardt et al., 2004).

Individuals orient attention in systematically biased ways, as a function of chronic and temporary factors. Moreover, they are unable to deploy attention to everything in their surroundings. These richly supported empirical findings undermine the assumption that all people experience the world as it is (for further review, see Granot et al., 2018). If perception did work in that manner, researchers would find that visual search patterns remain consistent across individuals, across time, and regardless of psychological factors including expectations, cognitive accessibility of other content, or context.

Unwarranted Confidence in Visual Experience

Though evidence suggests that individuals experience misperception and that divergent visual experiences point to potential inaccuracies at the level of the group, people tend to feel that their visual experiences represent reality completely and objectively (Feigenson & Spiesel, 2009; Griffin & Ross, 1991). Perhaps the most iconic example of people's trust in the visual system is exemplified by the first screening of the 1895 short film by Auguste Lumiere, "The Arrival of the Train," in which a camera positioned on the platform captures a train headed into a French station. The visual angle depicts the

train as if it is headed directly toward the viewer. At the premiere, audiences purportedly jumped from their seats in terror, believing that the train would burst through the screen (Barnouw, 1993; see Silbey, 2004, for a discussion). Despite knowing that they were not actually at a train station, their visual experience prompted genuine fear for their safety.

This steadfast belief in the evidence of one's eyes is known as *naïve realism*, or the belief that people see the world as it is, and can be understood in common adages such as "seeing is believing" that litter modern discourse (Ross & Ward, 1996). Axioms about trust in visual experiences are not new and range as far back as Aristotle, who reasoned, "of all the senses, trust only the sense of sight" (Aristotle, 350 BCE/1925). Indeed, some of the forefathers of psychological science advocated that human senses, and sight in particular, reflect the world as it really is. People see accurately, and as a result their understanding of the world as they see it is right (Griffin & Ross, 1991; Ross et al., 2010).

The tenets of naïve realism overlap meaningfully with those of perceptual dogmatism, demarcated by Tucker in this volume (see Chapter 7). Naïve realism is the idea that we see the world objectively—that is, that our perceptual experiences are unaffected by biases and, therefore, are true representations of what we lay our eyes on. Relatedly, *perceptual dogmatism* concerns the evidential value of perceptual experience, that is, the belief that we have evidence for what we think is true. In other words, naïve realism describes the belief that what people see is a true representation of what is there, and perceptual dogmatism states that perceptual experiences provide evidence to support that belief. Our goal is not to posit that a biased perceptual experience is necessarily an inaccurate perceptual experience per se. Rather, we argue that the justification of perceptual dogmatism and the unquestioning certainty of naïve realism may lead people to unwaveringly accept their visual experiences.

Indeed, the belief that one experiences the world as it truly is can have devastating consequences. In February of 1999, when plain-clothed New York City police officer Sean Carroll watched the 23-year-old, Black, Guinean immigrant Amadou Diallo pull a small, dark object out of his pocket while standing in the entryway of his apartment building under a lightbulb that had burned out, Carroll was certain that he saw a gun in the man's hand. He had prima facie justification for his belief and naively maintained that his visual experience reflected the only possible interpretation of what Diallo was holding. Carroll was wrong but did not know that before opening fire.

Diallo died, having been shot 41 times by Carroll and three additional officers who responded to Carroll's call, "Gun!"

Mitigating naïve realism in visual attention is a challenge because of both the certainty people hold in and the bias people experience with their own visual experiences. Vision confers a level of confidence that other forms of information do not. People are most confident in their interpretations when they are formed as a result of what they see for themselves, in contrast to what they learn through other sources of information. For example, participants either viewed real dashboard camera footage of a police–civilian altercation or read about the event from the perspective of the officer, the officer and the suspect, or an uninvolved eyewitness. While participants who strongly rather than weakly identified with police were more likely to evaluate the events in a manner that favored the officer, participants who watched the video were even more polarized as a function of their sense of identification when considering the certainty of their conclusions relative to those who received other forms of evidence (Sommers, 2016). Visual input, more so than other forms of input, can augment perceivers' confidence.

The problem is that confidence does not always track accuracy in the context of visual experience. In one study, participants were given the task of determining the direction of two motion stimuli presented simultaneously but in different positions on their screens. Researchers found that selective attention to a given target increased decision certainty by a significantly greater magnitude than the increase in accuracy (Zizlsperger et al., 2012). Similarly, in another task, participants indicated whether a patch of fuzzy lines tilted to the left or right, as well as their confidence in that judgment. The researchers tracked attention and found that patterns of selective attention predicted greater accuracy. However, participants' confidence was high regardless of where they directed attention (Wilimzig et al., 2008). This dissociation suggests that people need to encode very little visual information for their confidence in their perceptual experience to be high.

Even in the absence of visual input, such inflated confidence can arise. Because the optic nerve attaches to the retina, there are portions of each eye that do not contain rods or cones, resulting in blind spots from which no direct visual input can be obtained. The brain ensures that people do not perceive this gap, by filling in that spot in the visual field based on the surrounding information. Functionally, then, perceptual experience of information that falls on the blind spot is a sophisticated estimate and therefore

less reliable than directly perceived content. Despite this, people still had great confidence in their visual experience of information that researchers specifically projected onto the blind spot (Ehinger et al., 2017). In this way, not only do people fail to question their visual experience but they may confabulate missing visual information and feel just as confident in the veracity of their own conclusions.

Certain extreme physiological conditions can further demonstrate the dissociation between confidence and accuracy in visual perception. In the case of Anton's disease, patients suffer damage to the occipital lobes and are considered "cortically blind." Patients with Anton's, however, often still insist that they can see, despite contradictory indicators, like walking into obstacles (e.g., Roos et al., 1990). In addition, some patients experience blindsight, where they engage with their environment in ways that demonstrate accurate perception, despite their own articulated insistence of visual blindness; one famous patient, for instance, was blind in half of his visual field as a result of brain surgery decades earlier but still could guess with extreme accuracy when a stimulus appeared on a computer screen (Weiskrantz, 1986). Although these are extreme physiological conditions, the symptoms demonstrate that confidence, or even the lack thereof, is not always a good indicator of the accuracy of perceptual experience.

Confidence in visual experience, even despite relatively poor performance, is a challenge to mitigate because people feel that they themselves are not susceptible to misperception. Consider research on change blindness. This effect demonstrates that people can orient eye gaze directly on a given target while lacking the experience of consciously seeing it. Simons and Chabris (1999) found that 46% of observers watching a ball-passing game failed to notice a gorilla walking through the center of the scene, even when their eyes fixated on the costumed character (Memmert, 2006). In another context, 83% of highly trained radiologists also missed a gorilla, 48 times the size of an average nodule, embedded into an X-ray image of a pair of lungs, despite the fact that the majority of those who missed the gorilla looked directly at it, as confirmed by eye-tracking (Drew et al., 2013). Though inaccurate, people confidently overestimated their ability to detect such changes (Levin et al., 2002). For example, even after learning about the classic change blindness effect, including how and when the change occurred, participants underestimated their own susceptibility to misperception. While previous experiments had shown that 89% of people failed to notice key information, only 17% of participants acknowledged that they themselves could have

missed the change (Levin et al., 2000). Even in the face of countervailing evidence, people believe that they are not susceptible to error.

Moreover, people confidently assert that they are uniquely positioned to experience the visual world with a level of accuracy that others lack. For example, in legal cases trying negligence, radiologists supplied with knowledge about the outcome of a case sometimes testify against other radiologists for failing to detect tumors (Berlin, 2000). Researchers find that in 90% of cases, practitioners reported that a tumor was "visible in retrospect" (Muhm et al., 1983). Likewise, after viewing footage of altercations captured on police body-worn cameras, participants reported that they themselves were significantly more likely than the average American to objectively perceive and remember the events of the case (Jones et al., 2018). People believe themselves uniquely capable of perceiving the world as it is.

Confidence in the veracity of visual experience might be particularly difficult to curb because visual information is given primacy relative to other sources of input. The position of prominence ascribed to vision is reified in humans' biological construction. Neurons that process visual information occupy nearly 30% of the brain's cortex, compared to roughly 8% devoted to touch and 3% devoted to audition (Grady, 1993). People remember visual information better than auditory information. When individuals heard spoken information, the likelihood of remembering it in 3 days was 10%; when that same information was supplemented with a picture, the likelihood of accurate recall jumped to 65% (Medina, 2008). Further, visual processing increases the seeming veridicality of information. Semantic facts paired with a visual image were considered to be more truthful than those same statements without accompanying pictures (Newman et al., 2012). These factors suggest that naïve realism about visual evidence may be a unique case of naïve realism more generally (Feigenson & Spiesel, 2019).

Even in domains where other senses should dominate, vision is prioritized. Boston, Chicago, New York City, and Philadelphia use blind auditions when selecting membership in their prestigious symphony orchestras because of their understanding of the potentially misleading power of sight. Statisticians analyzed the likelihood of advancing through the audition phases for musicians who performed in both blind and non-blind auditions between 1970 and 1996. They found that a screen, occluding the performer from the eyes of evaluators, increased the probability that a woman would advance from the preliminary rounds by 50% and increased the likelihood that a woman would be selected in the final round by a factor of 7 (Goldin &

Rouse, 2000). Even in a domain predicated on sound, protecting what one hears from the influence of what one sees—even among those most highly skilled and trained to do so—is a challenge.

When input from our eyes competes with and actively contradicts input from other senses, we often prioritize what we see. In one study, researchers showed participants a miniature square through a visual distortion while simultaneously asking them to grasp it. What they saw conflicted with what they felt. But without realizing it, participants' estimates of the size of the square aligned better with their visual experience than their tactile one (Rock & Victor, 1964). In another experiment, researchers directed participants to hide one of their hands under a cover and in its place where their hand seemed like it should rest, participants saw a rubber hand matching their own. When a finger of the fake hand was bent back in a manner meant to appear painful, participants' skin conductance responses exhibited an intensity akin to real pain; the mere visual perception elicited physiological responses in the absence of any actual tactile sensation (Armel & Ramachandran, 2003). Despite knowing otherwise, what they saw looked, and therefore felt, real.

Naïve Realism and Biased Solicitation of Visual Information

Despite ample evidence of discrepancies in visual experience, naïve realism persists. People continue to think that what they see is unbiased and accurate. We argue that one reason for this is the fact that people solicit visual information in biased ways. People form hypotheses about what a scene will be comprised of, what an altercation will entail, and how an event will unfold. They selectively scan and devote processing resources toward visual elements that allow them to test and, more specifically, to confirm those hypotheses. Thus, people less frequently receive information that runs contrary to their beliefs and expectations about visual stimuli—a sort of *visual confirmation bias* (see Qu-Lee et al., in press). Just as people interpret information in line with their pre-existing beliefs (Nickerson, 1998) and weigh evidence more heavily that supports the conclusions they desire (Kunda, 1990), so too do people encode their visual surroundings and attend to their environments in ways that support their expectations (see Rajsic et al., 2017). For example, when scanning an image, people form a first impression of what the scene depicts. Their visual search patterns help inform those first impressions. When given

a second opportunity to view the scene, participants could seek out novel information to which they had not previously attended and in so doing gain a more complete understanding of what the scene entails. But they often do not do this. Instead, the locations on which participants fixate visual attention upon second viewing are more similar to their initial patterns than would be expected by chance (Noton & Stark, 1971; Underwood et al., 2009). Such evidence suggests that people seek out the same information multiple times as if forming a hypothesis upon first pass and seeking information to confirm it on the second.

Notably, expertise is insufficient to override visual confirmation bias. Five forensic experts with an average of 17 years of experience each and 85 years collectively demonstrated a similar susceptibility (Dror et al., 2006). Researchers provided these five experts with fingerprints left at crime scenes that the same five experts had identified in prior investigations as positively matching the suspects in question. The researchers included information in the case file that implied that the fingerprints would not match. Specifically, they told the experts that the two sets were from the Madrid bomber case, in which FBI agents erroneously matched the prints of an innocent man to those of the perpetrator. In this experiment, four out of the five experts changed their initial match decision, claiming that the fingerprints were definite non-matches or that there was insufficient information to form a decision. Decisions about the similarity of the prints reflected more than just the visual information provided. Irrelevant and in this particular case erroneous contextual information directed how the majority of these forensic experts saw the visual evidence.

Conclusions

In 1984, New York's Museum of Modern Art presented the *International Survey of Painting and Sculpture* exhibition. The show featured the works of 148 men but only 13 women and no artists of color. In response, Guerrilla Girls, a collective of female artists intent on exposing sexual and racial discrimination in art and culture, formed. The artists protected their identities by wearing gorilla masks in public. They worked under pseudonyms appropriated from deceased and remarkable women like Frida Kahlo and Gertrude Stein. Because exposure in major shows like those at leading cultural institutions increases the value of represented artists' work, Guerrilla

Girls created a poster campaign criticizing museums, dealers, curators, and critics they believed were complicit in the exclusion of women and non-White artists from social discourse and the financial boon the art world saw in the 1980s. One poster created by Guerrilla Girls stands out particularly because of what it does not include. Two-thirds of the visual space of this piece is blank. Nothing appears on the majority of the paper, but pressed far against the right edge are the words, printed in black block letters, "You're seeing less than half the picture."

As this protest makes clear, it is a misrepresentation—of art and of reality—when the mental image formed fails to include parts of what is really out there. But more than just the systematic exclusion of some from the public eye, Guerrilla Girls protested the lack of recognition of this bias. The art world failed to recognize its own blindness.

Far beyond the confines of high culture, people of all sorts hold a deep-seated faith in the reliability of their visual experiences, believing what they see to always be an accurate and complete representation of the world. The trust and confidence people have in visual experience can be particularly problematic in the court of law. The adversarial legal system is predicated on the idea that a jury of peers, each with potential biases they bring to bear in interpreting evidence, will as a whole reach a verdict approaching the truth as a result of an open deliberation processes (Swift, 2003). Yet the very principles of naïve realism undermine the basis for this tenet of due process. Not only do people consider their perceptions to be accurate and veridical, but they also consider their perceptions to be more objective than others' perceptions. People tend to underweight the input of others about the same information (Liberman et al., 2012). As a result, even deliberation among a jury of peers may be unlikely to compel people to consider or "see" new information or alternatives in visual evidence.

Moreover, the unwavering belief in one's own perceptual experiences coupled with the bias in those percepts that arise as a result of individuals' unique experiences can undermine faith in the legal system. The perceived and actual legitimacy of the legal system rest on verdicts aligning with the actual facts of the case. Legal scholars distinguish *substantive truth*—the actual facts of the case and reality of events—from *formal legal truth*—the facts as decided by jurists and the courts through legal proceedings (Summers, 1999). When factors shift how and to what individuals attend, the potential for greater divergence between substantive and formal truth emerges. As a result, the legitimate standing of the courts erodes. When ultimate decisions

rest on a dichotomous choice of guilt or innocence, even a minor bias in visual perception may hold the potential to tip the scales of justice.

References

Alonso, M., & Sutton, J. (2021). Ma'Khia Bryant was shot 4 times by officer, autopsy shows. *CNN*. Retrieved from https://www.cnn.com/2021/08/18/us/makhia-bryant-autopsy/index.html

Aristotle. (1925). *Metaphysics* (W. D. Ross, Trans.). Massachusetts Institute of Technology. http://classics.mit.edu/Aristotle/metaphysics.html (Original work published 350 BCE)

Armel, K. C., & Ramachandran, V. S. (2003). Projecting sensations to external objects: Evidence from skin conductance response. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(1523), 1499–1506.

Barr, A. (2017, November 21). FWPD moving to auto-record on police body cameras. NBC. https://www.nbcdfw.com/news/local/FWPD-Moving-to-Auto-Record-on-Police-Body-Cameras-459182863.html

Barnouw, E. (1993). *Documentary: A history of the non-fiction film*. New York: Oxford University Press.

Bauer, A., Schneider, S., Waldorf, M., Braks, K., Huber, T. J., Adolph, D, & Vocks, S. (2017). Selective visual attention towards oneself and associated state body satisfaction: An eye-tracking study in adolescents with different types of eating disorders. *Journal of Abnormal Child Psychology*, *45*(8), 1647–1661.

Berlin, L. (2000). Hindsight bias. *American Journal of Roentgenology*, *175*(3), 597–601.

Bosman, J., Robertson, C., Eckholm, E., & Oppel, R. A., Jr. (2014, November 25). Amid conflicting accounts, trusting Darren Wilson. *The New York Times*. https://www.nytimes.com/2014/11/26/us/ferguson-grand-jury-weighed-mass-of-evidence-much-of-it-conflicting.html

Bouhuys, A. L., Geerts, E., & Gordijn, M. C. M. (1999). Depressed patients' perceptions of facial emotions in depressed and remitted states are associated with relapse: A longitudinal study. *Journal of Nervous and Mental Disease*, *187*(10), 595–602.

Brady, W., & Balcetis, E. (2015). Accuracy and bias in emotion perception predict affective response to relationship conflict. In T. Heinin (Ed.), *Advances in visual perception research* (pp. 29–43). Nova Science Publishers.

Capehart, J. (2015, July 16). How Eric Garner changed the national conversation on race and police. *The Washington Post*. https://www.washingtonpost.com/blogs/post-partisan/wp/2015/07/16/how-eric-garner-changed-the-national-conversation-on-race-and-police/?utm_term=.89cb25297817

Carroll, M. (2014). New York City voters want their broken windows fixed, Quinnipiac University poll finds; "No excuse" for Garner death, voters say almost 3–1. Quinnipiac University. https://newyork.cbslocal.com/wp-content/uploads/sites/14578484/2014/08/quinnipiac.pdf

CBS News/New York Times. (2015). Poll April 30–May 3, 2015. http://assets.documentcloud.org/documents/2072085/apr15d-race-trn.pdf

Cineas, F. (2021). Why they're not saying Ma'Khia Bryant's name. *Vox*. Retrieved from https://www.vox.com/22406055/makhia-bryant-police-shooting-columbus-ohio

Donders, N. C., Correll, J., & Wittenbrink, B. (2008). Danger stereotypes predict racially biased attentional allocation. *Journal of Experimental Social Psychology*, *44*, 1328–1333.

Drew, T., Võ, M L. H., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science*, *24*(9), 1848–1853.

Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, *156*(1), 74–78.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing Black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, 87(6), 876–893.

Editorial Board. (2017, July 22). Officers, turn on your body cameras. *The Washington Post*. https://www.washingtonpost.com/opinions/officers-turn-on-your-body-cameras/2017/07/22/41290ff0-6e3e-11e7-b9e2-2056e768a7e5_story.html

Ehinger, B. V., Häusser, K., Ossandon, J. P., & König, P. (2017). Humans treat unreliable filled-in percepts as more real than veridical ones. *eLife*, 6, Article e21761. https://doi.org/10.7554/eLife.21761

Farid, H. (2009). Seeing is not believing. *IEEE Spectrum*, *46*(8), 42–47.

Feigenson, N., & Spiesel, C. (2009). *Law on display: The digital transformation of legal persuasion and judgment*. New York University Press.

Feigenson, N., & Spiesel, C. (2019). The psychology of surveillance and sousveillance video evidence. In C. J. Najdowski & M. C. Stevenson (Eds.), *Criminal juries in the 21st century: Contemporary issues, psychological science, and the law* (pp. 173–193). Oxford University Press.

Ferguson and Other Cities React to Grand Jury Decision not to Indict Darren Wilson. (2014, November 24). *The New York Times*. https://news.blogs.nytimes.com/2014/11/24/live-updates-from-ferguson-on-the-grand-jury-decision-in-michael-brown-shooting/

Floyd v. City of New York. 959 F. Supp. 2d 540 (2013).

Foley, A., Foley, H. J., Scheye, R., & Bonacci, A. (2007). Remembering more than meets the eye: A study of memory confusions about incomplete visual information. *Memory*, *15*(6), 616–633.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, *90*(4), 715–741.

Goodwin, H., Eagleson, C., Mathews, A., Yiend, J., & Hirsch, C. (2017). Automaticity of attentional bias to threat in high and low worriers. *Cognitive Therapy and Research*, *41*(3), 479–488.

Gotlib, I. H., Krasnoperova, E., Yue, D. N., & Joorman, J. (2004). Attentional biases for negative interpersonal stimuli in clinical depression. *Journal of Abnormal Psychology*, *113*(1), 127–135.

Grady, D. (1993, June 1). The vision thing: Mainly in the brain. *Discover Magazine*. http://discovermagazine.com/1993/jun/thevisionthingma227

Granot, Y., Balcetis, E., Feigenson, N., & Tyler, T. R. (2018). In the eyes of the law: Perception versus reality in appraisals of video evidence. *Psychology, Public Policy, and Law*, 24, 93–104.

Griffin, D. W., & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. *Advances in Experimental Social Psychology*, *24*, 319–359.

Isaacowitz, D. M., Löckenhoff, C. E., Lane, R. D., Wright, R., Sechrest, L., Riedel, R., & Costa, P. T. (2007). Age differences in recognition of emotion in lexical stimuli and facial expressions. *Psychology and Aging*, *22*(1), 147–159.

Jang, S.-K., Kim, S., Kim, C.-Y., Lee, H.-S., & Choi, K.-H. (2016). Attentional processing of emotional faces in schizophrenia: Evidence from eye tracking. *Journal of Abnormal Psychology*, *125*(7), 894–906.

Jones, K. A., Crozier, W. E., & Strange, D. (2018). Objectivity is a myth for you but not for me or police: A bias blind spot for viewing and remembering criminal events. *Psychology, Public Policy, and Law*, *24*(2), 259–270.

Jones, K. A., Crozier, W. E., & Strange, D. (2019). Look there! The effect of perspective, attention, and instructions on how people understand recorded police encounters. *Behavioral Sciences & the Law*, *37*(6), 711–731.

Joormann, J., & Gotlib, I. H. (2006). Is this happiness I see? Biases in the identification of emotional facial expressions in depression and social phobia. *Journal of Abnormal Psychology*, *115*(4), 705–714.

Kahan, D. M., Hoffman, D. A., & Braman, D. (2009). Whose eyes are you going to believe? Scott v. Harris and the perils of cognitive illiberalism. *Harvard Law Review*, *122*(3), 837–906.

Kassin, S. M. (2002, November 1). False confessions and the jogger case. *The New York Times*. http://www.nytimes.com/2002/11/01/opinion/false-confessions-and-the-jogger-case.html

Kassin, S. M., Drizin, S. A., Grisso, T., Gudjonsson, G. H., Leo, R. A., & Redlich, A. D. (2010). Police-induced confessions: Risk factors and recommendations. *Law and Human Behavior*, *34*, 3–38.

Kaye, D. H., Broun, K. S., Dix, G. E., Imwinkelried, E. J., Mosteller, R. P., Roberts, E. F., & Swift, E. (2013). *McCormick on evidence* (7th ed.). Thomson Reuters.

Kenny, D. A., & Acitelli, L. K. (2001). Accuracy and bias in the perception of the partner in a close relationship. *Journal of Personality and Social Psychology*, *80*(3), 439–448.

Kruglanski, A. W. (1989). The psychology of being "right": The problem of accuracy in social perception and cognition. *Psychological Bulletin*, *106*(3), 395–409.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Levin, D. T., Drivdahl, S. B., Momen, N., & Beck, M. R. (2002). False predictions about the detectability of visual changes: The role of beliefs about attention, memory, and the continuity of attended objects in causing change blindness blindness. *Consciousness and Cognition*, *11*(4), 507–527.

Levin, D. T., Momen, N., Drivdahl, S. B., & Simons, D. J. (2000). Change blindness blindness: The metacognitive error of overestimating change-detection ability. *Visual Cognition*, *7*(1–3), 397–412.

Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the wisdom of dyads. *Journal of Experimental Social Psychology*, *48*(2), 507–512.

Lukáts, A., Szabó, A., Röhlich, P., Vígh, B., & Szél, Á. (2005). Photopigment coexpression in mammals: Comparative and developmental aspects. *Histology and Histopathology*, *20*(2), 551–574.

Mansfield, M., & Kenton, L. (2021). 'Knife wielding maniac' Candace Owens says Ma'Khia Bryant, 16, was trying to 'butcher another human' when she was 'killed by cop.' *The US Sun*. Retrieved from https://www.the-sun.com/news/2746679/candace-owens-brands-makhia-bryant-knife-wielding-maniac/

Marquart, F., Matthes, J., & Rapp, E. (2016). Selective exposure in the context of political advertising: A behavioural approach using eye-tracking methodology. *International Journal of Communication*, *10*, 2576–2595.

Medina, J. (2008). *Brain rules: 12 principles for surviving and thriving at work, home, and school*. Pear Press.

Memmert, D. (2006). The effects of eye movements, age, and expertise on inattentional blindness. *Consciousness and Cognition*, *15*(3), 620–627.

Morris, E. (2011). *Believing is seeing (observations on the mysteries of photography)*. Penguin Books.

Morris, F. (2021). Columbis activists call for federal probe of police after Ma'Khia Bryant shooting. *NPR*. Retrieved from https://www.npr.org/2021/04/22/990029491/columbus-activists-call-for-federal-probe-of-police-after-makhia-bryant-shooting

Muhm, J. R., Miller, W. E., Fontana, R. S., Sanderson, D. R., & Uhlenhopp, M. A. (1983). Lung cancer detected during a screening program using four-month chest radiographs. *Radiology*, *148*(3), 609–615.

Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., & Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review*, *19*(5), 969–974.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Niedenthal, P. M., Brauer, M., Robin, L., & Innes-Ker, A. H. (2002). Adult attachment and the perception of facial expression of emotion. *Journal of Personality and Social Psychology*, *82*(3), 419–433.

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern recognition. *Science*, *171*(3968), 308–311.

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, *130*(3), 466–478.

Pearce, M., Kaleem, J., Etehad, M., & Winton, R. (2017, October 12). In Las Vegas, the casino is always watching—And yet it missed Stephen Paddock. *LA Times*. http://www.latimes.com/nation/la-na-vegas-shooting-casino-security-20171012-story.html

Pew Research Center. (2014, December 8). Sharp racial divisions in reactions to Brown, Garner decisions. http://www.people-press.org/2014/12/08/sharp-racial-divisions-in-reactions-to-brown-garner-decisions/

Phippen, J. W. (2016, August 2). The trouble with police body cameras. *The Atlantic*. https://www.theatlantic.com/news/archive/2016/08/chicago-shooting/493997/

Qu-Lee, J., Seidel, B., Harel, D., Granot, Y., & Balcetis, Y. (in press). The relationship between visual confirmation, belief consistency, and belief polarization. *Comprehensive Results in Social Psychology.*

Rajsic, J., Taylor, J. E. T., and Pratt, J. (2017). Out of sight, out of mind: Matching bias underlies confirmatory visual search. *Attention, Perception, & Psychophysics*, *79*(2), 498–507.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rock, I., & Victor, J. (1964). Vision and touch: An experimentally created conflict between the two senses. *Science*, *143*(3606), 594–596.

Roefs, A., Jansen, A., Moresi, S., Willems, P., van Grootel, S., & van der Borgh, A. (2008). Looking good. BMI, attractiveness bias and visual attention. *Appetite*, *51*, 552–555.

Roos, K. L., Tuite, P. J., Below, M. E., & Pascuzzi, R. M. (1990). Reversible cortical blindness (Anton's syndrome) associated with bilateral occipital EEG abnormalities. *Clinical Electroencephalography*, *21*(2), 104–109.

Ross, L., Lepper, M., & Ward, A. (2010). History of social psychology: Insights, challenges, and contributions to theory and application. In S. T. Fiske, D. T. Gilbert, & L. Gardner (Eds.), *Handbook of social psychology* (Vol. 1). Wiley. https://doi.org/10.1002/9780470561119.socpsy001001

Ross, L., & Ward, A. (1996). Naïve realism in everyday life: Implications for social conflict and misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *Values and Knowledge* (pp. 103–135). Hillsdale, NJ: Erlbaum.

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination Cognition and Personality*, *9*, 185–211.

Sanburn, J. (2014, November 26). The one battle Michael Brown's family will win. *Time*. http://time.com/3606376/police-cameras-ferguson-evidence/

Scott v. Harris, 127 S. Ct. 1769, 1781 (2007).

Silbey, J. M. (2004). Judges as film critics: New approaches to filmic evidence. *University of Michigan Journal of Law Reform*, *37*(2), 493.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*, 1059–1074.

Sommers, R. (2016). Will putting cameras on police reduce polarization. *Yale Law Journal*, *125*(5), 1304–1362.

Stephan, W. G., & Stephan, C. W. (2000). An integrated threat theory of prejudice. In *The Claremont Symposium on applied social psychology: Reducing prejudice and discrimination* (pp. 23–45). Lawrence Erlbaum Associates.

Sternisko, A., Granot, Y., & Balcetis, E. (2017). One-sighted: How visual attention biases legal decision-making. In D. Delagarza (Ed.), *New developments in visual attention research* (pp. 105–139). Nova.

Stoughton, S. W. (2018). Police body-worn cameras. *North Carolina Law Review*, *96*, 1363–1424.

Summers, R. (1999). Formal legal truth and substantive truth in judicial fact-finding: Their justified divergence in some particular cases. *Law and Philosophy*, *18*, 497–511.

Swift, E. (2003). Aspirational optimism about evidence law: An implicit theme of the *Visions of Rationality* symposium. *Michigan State Law Review*, 4, 1337–1364.

Treisman, A. (2006). How the deployment of visual attention determines what we see. *Visual Cognition*, *14*(4–8), 411–443.

Turner, B. L., Caruso, E. M., Dilich, M. A., & Roese, N. J. (2019). Body camera footage leads to lower judgments of intent than dash camera footage. *Proceedings of the National Academy of Sciences*, *116*(4), 1201–1206.

Underwood, G., Foulsham, T., & Humphrey, K. (2009). Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition. *Visual Cognition*, *17*(6/7), 812–834.

Wadlinger, H. A., & Isaacowitz, D. M. (2006). Positive mood broadens visual attention to positive stimuli. *Motivation and Emotion*, *30*(1), 87–99.

Ware, L. J., Lassiter, G. D., Patterson, S. M., & Ransom, M. R. (2008). Camera perspective bias in videotaped confessions: Evidence that visual attention is a mediator. *Journal of Experimental Psychology: Applied*, *14*(2), 192–200.

Weiskrantz, L. (1986). *Blindsight: A case study and implications*. Oxford University Press.

Wilimzig, C., Tsuchiya, N., Fahle, M., Einhäuser, W., & Koch, C. (2008). Spatial attention increases performance but not subjective confidence in a discrimination task. *Journal of Vision*, *8*(7), 1–10.

Zaki, J., Weber, J., Bolger, N., & Ochsner, K. (2009). The neural bases of empathic accuracy. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(27), 11382–11387.

Zizlsperger, L., Sauvigny, T., & Haarmeier, T. (2012). Selective attention increases choice certainty in human decision making. *PLoS One*, *7*(7), Article e41136.

7

Dogmatism and the Epistemology of Covert Selection

Chris Tucker

You and I are walking down the street. You look where you are going—straight ahead—because you don't want to run into anyone. My desire for dessert causes me to look left at the window of the pastry shop. Here my desire's influence on experience is *overt cognitive selection*. It is *selection* insofar as my desire influences the character and content of my experience by influencing my attention. It is *cognitive* insofar as the state doing the influencing, namely the desire, is a cognitive state. In philosophy, it is typical to count at least the following things as cognitive states: beliefs, desires, moods, experiences, emotions, preferences, expectations, and concept possession. The influence is *overt* insofar as the desire's mediate influence on attention occurs by influencing the position and orientation of my body, especially my sensory organs.

I now look ahead, eyes front and center, but my desire for dessert continues to influence my perceptual experience. In my peripheral vision (and with my eyes facing straight ahead), I continue to linger on those marvelous pastries. My desire's influence is now an instance of *covert selection*. It is *covert* insofar as the desire's mediate influence on attention does not occur by influencing the position and orientation of my body, including my sensory organs.

My nearly overwhelming desire for dessert has one final influence on my experience before I distractedly walk into a pole. While lingering on the pastries in my peripheral vision, my desire for pastry causes my experience to represent those pastries as closer than they actually are. Here the desire's influence—if such influence is even compatible with our actual hardwiring—counts as *cognitive penetration*. Again, the desire's influence would be *cognitive* because a cognitive state is having the influence. It is *penetration* insofar as the desire's influence would be, in some hard to specify sense, direct.

Chris Tucker, *Dogmatism and the Epistemology of Covert Selection* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press.
DOI: 10.1093/oso/9780197636916.003.0008

To be direct is to be unmediated. If our desires influence our experience at all, there is presumably some sort of processing that mediates the influence our desires have on our experience. For it to be an interesting question of whether our cognitive states penetrate our experience, the relevant sort of directness must be compatible with some kinds of mediation; otherwise, it will be very easy to show that our cognitive states don't penetrate our experience.

Influence mediated by overt selection is uncontroversially indirect (see Gross, 2017, p. 3); insofar as some influence is overt (cognitive[1]) selection, it is not penetration. It is controversial, however, whether influence mediated by covert selection is direct in the relevant sense, and thus it isn't clear what the relationship is between penetration and covert selection. The disparate accounts of covert attention's (and so covert selection's) ontology only make the relationship more obscure.[2]

Suppose that your perceptual experience represents that the Black guy is holding a gun, where the experience results from prejudiced overt selection, covert selection, or penetration. How, if at all, would these prejudicial influences affect whether it is rational to believe what your experience tells you? With respect to evaluating the rationality of believing what your experience tells you, does it matter which of the three kinds of influence is at issue?

This chapter explores these questions by considering the implications of covert selection for a controversial but popular position in epistemology, misleadingly labeled "perceptual dogmatism." *Perceptual dogmatism* holds, roughly, that a perceptual experience is always evidence that its representational content is true. If your perceptual experience represents that the Black guy is holding a gun, then the experience is evidence that the Black guy is holding a gun. Indeed, perceptual dogmatism holds that it is evidence that he is holding a gun no matter how it is caused. If cognitive selection or penetration of an experience can make a difference to whether the experience is evidence, then perceptual dogmatism is false.

Overt selection (in the guise of bad searches for evidence) and cognitive penetration pose well-known challenges to perceptual dogmatism.[3] The myriad accounts of covert attention's metaphysics make one (at least, they

[1] I generally suppress the *cognitive* qualifier from here on out.

[2] Attention has been claimed to be, for example, a subpersonal capacity or mechanism, a certain kind of subpersonal process, a way processing happens, a feature of experience, and a person-level relation between a thing and object.

[3] Traditionally, the bad search challenges target most directly not perceptual dogmatism but the thesis that undefeated evidence suffices for justification. But the challenge extends equally to perceptual dogmatism.

made me) wonder whether it raises a distinctive challenge to dogmatism, one that is importantly different from the challenges that dogmatism already faces. Epistemological issues concerning covert selection cannot be entirely divorced from the metaphysical ones (e.g., see below, "No New Direct Challenge"); however, I argue that no matter how the metaphysical issues get sorted, covert selection fails to provide a distinctive challenge to dogmatism. This chapter is good news for dogmatists. The fewer distinct challenges to dogmatism there are, the more likely a dogmatist can resolve them all.

For those who couldn't care less about dogmatism, you'll be relieved to hear that what I have to say is of broader interest. For psychologists, perceptual dogmatism is a natural place to begin thinking about epistemology. The view is simple and intuitive, and yet it is connected to many major disputes in epistemology. By exploring how dogmatism can be challenged, you will be introduced to the field of epistemology. For philosophers, the epistemology of covert attention is a relatively new area with no established positions or overarching framework. A clear exposition of how covert attention is related (or not) to a prominent position in epistemology is a natural place to begin understanding the epistemic significance of covert selection. By the end of the chapter, we will have considered a variety of ways in which covert selection might be epistemically significant.

In the second section, I explain (perceptual) dogmatism and the jargon needed to state it more precisely. In the third section, I explain how cognitive penetration raises a significant challenge to dogmatism. In the fourth section, I explain how overt selection raises a distinct, significant challenge to dogmatism. The goal in these early sections is to identify a working idea of how covert selection would have to challenge dogmatism if it is to be importantly different from dogmatism's other challenges.

The remaining two sections of the chapter admit that covert selection can raise various challenges for dogmatism, but they contend that the challenges are ones that dogmatism already faces. In the fifth section, I argue that if covert selection poses a direct challenge to dogmatism's truth (i.e., if it provides evidence that dogmatism is false), it will collapse into the challenge posed by either cognitive penetration or overt selection. In the sixth section, I consider whether covert selection might raise a new worry for an explanatory ambition of typical dogmatists. If so, then covert selection would pose an indirect challenge for dogmatism, removing a common motivation for the view. I concede that covert selection can raise such a challenge, but, even here, the challenge is not new.

Perceptual Dogmatism

Defining Perceptual Dogmatism

Epistemologists like to annoy people, so they invented a technical phrase, "X has justification to believe," that is roughly equivalent to more everyday locutions, such as "it is reasonable for X to believe" and "it is rational for X to believe." So understood, justification is a less demanding status than knowledge. It is usually assumed, for example, that you can have justification to believe something false but that you can't know something false. You might justifiably, or reasonably, believe that Chuck died in a plane crash on the basis of reliable news reports and footage of the crash; however, if he miraculously survived, then you don't know that he died in the plane crash.

Some epistemologists like to show off their Latin, so they invented another technical term, *prima facie justification*. For something to be a prima facie justification to believe P is, roughly, for it to be (a piece of) evidence that supports believing P. Some prima facie justifications are stronger than others because some pieces of evidence are stronger than others. And sometimes the counterevidence is stronger than the evidence. In such a case, we say that our prima facie justification is *defeated*. Mike's fingerprints on the bloody knife may be evidence that he did it; however, such evidence might be defeated by a massive body of counterevidence, including surveillance videos, wiretapped conversations, eyewitness testimony, and Ike's guilty confession that he framed Mike because Mike "stole" his girlfriend.

Prima facie justification and (unqualified) justification are closely linked. The "prima facie" functions as a qualifier, or a caveat. You have justification to believe something when it is prima facie justified and one further condition is met: You have no (sufficiently strong) defeaters or counterevidence.[4] Prima facie justification is evidence; justification is undefeated evidence.

There was lots of evidence for Newtonian physics' truth, and arguably before Einstein this evidence justified people in believing that Newtonian physics is true. Now, however, there is lots of evidence that Newtonian physics is strictly false no matter how well it works as an approximation. Before Einstein, physicists had both prima facie justification and (unqualified)

[4] There are at least two kinds of counterevidence, rebutting and undermining. Suppose you have some evidence for proposition P. Rebutting counterevidence is evidence that P is false. In contrast, undermining evidence might be evidence not that P is false but that your evidence for P is unreliable.

justification for believing Newtonian physics; afterward, they have prima facie justification but not (unqualified) justification to believe it. Since they now lack justification to believe it, it is ordinarily assumed that it would be unreasonable or irrational for them to believe it.

This chapter focuses on a particular claim about the relationship between perceptual experiences and what we have justification to believe. Suppose you have a perceptual experience that represents as true that there is a book in front of you, and suppose that the experience is the only potentially relevant consideration you have concerning whether P is true. Should you disbelieve that there is a book in front of you or withhold judgment (i.e., resist both belief and disbelief) about the matter? Presumably not. After all, the only potentially relevant consideration bearing on the matter is that you seem to see a book! You apparently have prima facie justification, or evidence, to believe that the book is there. And since we've stipulated that you have no counterevidence, you are justified in believing that the book is there.

There is nothing special about experiences of books or special about your experiences. We can generalize. The reasoning[5] from the previous paragraph leads to the following general theory about the relation between perceptual experience and prima facie justification:

> **(Perceptual) Dogmatism:** Necessarily, if S has a perceptual experience that P, then S has prima facie justification for believing P.

In other words, your perceptual experience that P is evidence that P is true, and as long as you have no relevant defeaters/counterevidence, it is reasonable to believe that P is true. Since I want you to like this view, it is counterproductive to refer to it as *dogmatism*. Who wants to endorse anything associated with dogmatism? I didn't coin the term. Just keep in mind that it doesn't refer to a stubborn adherence to some doctrine; instead, it refers to a widely endorsed thesis concerning the relation between perceptual experience and prima facie justification.[6]

[5] As attractive as I find this reasoning, no one has been able to turn it into a successful argument (see Tucker [in press; 2013, pp. 9–12] for an explanation of the problem).

[6] See, for example, Huemer (2001, 2013a), Pryor (2000), and Tucker (2010, 2014).

Perceptual Dogmatism versus Naïve Realism

In Chapter 6 of this volume, Granot, Jones, and Balcetis attack a position they call "naïve realism." Since both perceptual dogmatism and naïve realism are theses about perceptual experience, readers of this volume may wonder about their relationship. In the rest of this section, I explain naïve realism, how an attack on naïve realism might motivate an objection to perceptual dogmatism, and why Granot et al.'s attack fails.

Naïve realism has two components: "our perceptual experiences are unaffected by biases and, therefore, are true representations of what we lay our eyes on" (see Chapter 6). The first is a claim about a perceptual experience's causal history: The causal processes that produce perceptual experience do not systematically privilege any information with regard to how quickly it was processed, how frequently it was processed, or how intensely it valences an emotional display (e.g., moderately happy vs. very happy). The second is a claim about the accuracy of perceptual experience's content: All contents of the experience are accurate.[7]

Naïve realism, then, is a claim about the causal history and accuracy of perceptual experience. It is descriptive rather than evaluative. To be sure, the term *bias* is often used as a negative evaluation. But that is not the way that Granot et al. characterize it in their chapter. Whether perceptual processing is biased, as they characterize it, is just the question of whether certain inputs get certain kinds of priority in the perceptual processing. It is a separate question whether, for example, it is good or rational for these inputs to have these kinds of priority.

Perceptual dogmatism, in contrast, is an evaluative claim. In saying that perceptual experiences count as evidence no matter what, perceptual dogmatism is saying that it is rational (a positive evaluation) to trust them in the absence of countervailing considerations. I will make the relatively uncontroversial assumption that the accuracy of a perceptual experience—the truth/falsity of naïve realism's second component—makes no difference to whether an experience counts as evidence. In other words, it can be rational to rely on inaccurate perceptual experiences. Your perceptual experience

[7] They explicitly say that "Perceptual *accuracy* is defined as the ability to correctly identify the visual experience" (see Chapter 6). Here, I think they are using *ability* and *visual experience* too loosely. Based on the examples, I think they mean the following: Perceptual accuracy is the extent to which a given perceptual state (e.g., perceptual experience, perceptual judgment) correctly represents the target distal stimuli.

misrepresents the Mueller-Lyer display such that one horizontal line is represented as longer than the other. Before you discovered that your experience is illusory, it was presumably rational for you to believe that one line is longer than the other. (Also, see the discussion of the mad scientist cases in the next section.)

Perceptual dogmatism claims that any bias in the causal history is irrelevant to the rationality of trusting your experience—except insofar as those biases affect what kinds of counterevidence you might have. In other words, perceptual dogmatism claims that naïve realism's first component is also irrelevant to the rationality of trusting your experience. Here, critics of dogmatism disagree. If these critics are correct, then any evidence against naïve realism's first component may underwrite the following sort of objection to dogmatism: (1) There actually are various kinds of cognitive penetration and/or selection (as demonstrated by the alleged empirical evidence); (2) it isn't rational to trust experience when it is penetrated and/or selected in those ways (as the critics may insist); therefore, (3) it isn't always rational to trust your experience and, thus, perceptual dogmatism is false.

There's good news and bad news for Granot et al. The good news is that the problems with their attack on naïve realism do not afflict the broader goals of their chapter. They provide an impressively diverse range of evidence for the claim that perceptual judgments[8] are the result of bias and, consequently, are sometimes inaccurate and/or misleading. Potential biases in the perceptual judgments of judges and juries would be enough to show that video evidence is not the panacea it is often made out to be. Conscientious legal judgments made on the basis of video evidence may very well be biased, inaccurate, and unjust. It is to Granot et al.'s credit that they warn us of this danger.

The bad news is that there is a gap in Granot et al.'s arguments against naïve realism. Their empirical evidence generally concerns which perceptual judgments subjects make and which they don't make.[9] Yet, naïve realism, as they defined it, is a claim about perceptual experience.

[8] Here, I use *perceptual judgment* loosely. In the cases they discuss, it isn't always clear whether the biased judgment is solely based on the perceptual experience or whether it is closer to the conclusion of an inference from both perceptual and background beliefs. For an empirically informed discussion of which beliefs strictly count as perceptual, see Lyons (2009, Chapter 4).

[9] The vagueness of naïve realism is an additional problem. Naïve realism will be an implausible thesis, and so trivial to argue against, if it is the universal generalization that all perceptual experiences are accurate and unaffected by bias. Perhaps Granot et al. intend naïve realism as some sort of generic generalization. Yet true generics are compatible with "exceptions that prove the rule": Dogs have four legs, but three-legged dogs exist. Granot et al. could be clearer about why the alleged exceptions to naïve realism count as counterexamples rather than exceptions that prove the rule.

I take it that Granot et al. make two assumptions throughout the chapter, which are supposed to close the gap. First, they apparently assume that, in the cases that they discuss, inaccuracy or bias at the level of judgment is best explained by a matching inaccuracy or bias at the level of perceptual experience. The first assumption seems very plausible in some cases and questionable in others. Consider the fingerprint identification example in which background information made a difference to whether the fingerprints were judged to be a match. Granot et al. assume that this difference in judgment is due to a difference in what the experts saw. But why is that assumption better than the distinct empirical hypothesis that the difference in judgment is due not to what was seen but to what was inferred? They don't say.

The second assumption is that a visual experience misrepresents whenever it fails to represent everything that is in one's visual field or "parts of what is really out there." This assumption is questionable in most, if not all, the cases they have in mind. Suppose I say, "Obama was a US president," but I don't mention that he was indeed the first Black US president. What I said was accurate as far as it went, even if my failure to say more was somehow misleading in a given context. Suppose that a cop's perceptual experience represents a Black guy as having something in his hand, but the experience does not represent what is in the hand. If the Black guy is holding a cell phone, the experience did not thereby misrepresent the Black guy, what he was holding, or the visual scene. The experience was accurate as far as it went, even if it misleads the cop and tragic consequences follow. Due to the processing limitations mentioned by Granot et al., one might expect that we are consciously experiencing only a (biased) sample of what there is to experience in the scene. And if our experience isn't representing some aspect of the scene, it can't misrepresent that aspect either.[10]

[10] Granot et al. defend the second assumption by appealing to philosophical argument, namely an analogy with the *International Survey of Painting and Sculpture* exhibition. They claim that the show misrepresents something about the artists who make the best art (they don't specify what) because the show featured the work of only 148 men, 13 women, and no artists of color. We can agree that the exhibition problematically excluded minorities and may mislead people into thinking that art from minorities is not worth their consideration. Yet the exhibition, strictly speaking, misrepresents something about X only if it represents something about X. While it is at least somewhat plausible that the exhibition was representing—and misrepresenting—something about the artists who make the best art, it is implausible that it is representing every part "of what is really out there" or that it is some "complete representation of the world." The exhibition did not represent—and so did not misrepresent—anything about how darkly roasted I like my coffee (I like it no darker than medium). My perceptual experience misrepresents some part of my visual scene only if it is both representing that part of the visual scene and doing so inaccurately. A perceptual experience that doesn't represent the gorilla in the middle of the scene or doesn't represent the gorilla as a gorilla may nonetheless be accurate as far as it goes.

To be clear, I am not claiming that naïve realism is true. I am claiming only that Granot et al.'s attack on naïve realism fails because they did not justify the leap from *evidence concerning what judgments subjects (don't) make* to *conclusions concerning what subjects' perceptual experiences (don't) represent*. It is unclear, then, whether they've provided much empirical evidence that perceptual experience (as opposed to judgment) is cognitively penetrated and/or selected. Yet, as we'll see in the next section, perceptual dogmatism isn't off the hook if naïve realism is actually true. The mere possibility that biases can affect our perceptual experiences is enough to raise interesting challenges to dogmatism.

The Cognitive Penetration Challenge

The Challenge's Distinctive Issue

Philosophy is more fun than science. Science requires tedious experiments and lots of waiting around for the results. Note the term *necessarily* at the beginning of dogmatism's definition. Typical philosophical theories are such that, if they are true, they are true necessarily (i.e., no matter what). No matter whether human beings exist. No matter what the laws of nature are. No matter what. And this means that we can gain philosophical understanding just by thinking about stories or watching movies or imagining wild scenarios. At least, thinking about such things can lead to philosophical understanding so long as the story is possible. Because if it's possible and a philosophical theory is incorrect in the story, then it's incorrect period. If the theory is true, then it's true no matter what.

Science may have a central place in philosophy. Perhaps the leading scientific theories should constrain philosophical theorizing in various ways. Yet insofar as philosophy is concerned with what is true no matter what, it must go beyond what science tells us about the way things are—even beyond what science tells us about the laws of nature. As we assess the case for dogmatism, then, we needn't limit ourselves to what science has demonstrated about the actual causes and contents of perceptual experience.

Dogmatism is an equal opportunity employer: It doesn't discriminate perceptual experiences on the basis of their causal history. In the absence of relevant counterevidence, it says that a perceptual experience provides justification to believe its content no matter how it is caused. In some cases, this

looks like the correct result. Suppose that, unbeknown to you, a mad scientist uses a computer to generate your perceptual experiences so flawlessly that they feel genuine (e.g., you are in the Matrix or an *Inception* dream). This is not a good causal history for your experience to have, but as long as you (reasonably) have no idea that they are computer-generated, it is widely assumed that you have justification to believe what your experiences tell you. In fact, the mad scientist tricks you by capitalizing on your rationality (Kelly, 2014, section 2). He knows that reasonable people believe what their perceptual experiences tell them unless they have a good reason not to.

In other cases, dogmatism's policy of anti-discrimination looks mistaken. Suppose that Wishful Willy's perceptual experience represents the rock as a gold nugget only because he is overwhelmed by a desire to be rich. Or suppose that Jill has an irrational belief that Jack is angry and that this belief causes her experience to represent Jack's neutral expression as angry. Even if Willy and Jill have no idea that their cognitive states are influencing their perceptual experience, it seems irrational for Willy and Jill to believe what their experience tells them. These cases seem to be counterexamples to dogmatism.

If you find it hard to believe that Willy and Jill could have no idea that their experiences were influenced by their desire and fear, respectively, then so much the better for dogmatism. We may not get a counterexample at all. For then we can explain why they lack justification to believe the content of their experience by appealing to a defeater, namely their awareness of the experiences' poor causal history. For the sake of the chapter, we can just play along with the assumptions needed by those objecting to dogmatism.

Since the Willy and Jill cases are generally assumed to involve cognitive penetration, we'll call this challenge to dogmatism the *cognitive penetration challenge*.[11] This challenge is even more interesting if cognitive penetration is compatible with the hardwiring human beings actually have; however, its force as an objection to dogmatism requires only that such cases be possible.

The ultimate goal is to show that covert selection doesn't raise any new challenge to dogmatism that dogmatism doesn't already face. I'll err on the side of giving overly narrow construals of the distinctive issues raised by the

[11] Proponents of this challenge include Markie (2005, pp. 356–357; 2013), McGrath (2013), and Siegel (2012). Defenses of dogmatism from this challenge include Huemer (2013a, pp. 343–345; 2013b), Skene (2013, section 5.1), and Tucker (2010, section 6; 2014). Keep in mind that virtually every traditional rival to dogmatism, including reliabilism, faces a cognitive penetration challenge that is at least as bad as the one faced by dogmatism (see Tucker, 2014). So most epistemologists issue this challenge against dogmatism at their own peril.

existing challenges. This will make my job harder by making it easier for covert selection to raise a distinct challenge.

We criticize inferences for a wide variety of reasons, including that one lacks adequate justification to believe the premises, the inference amounts to a hasty generalization, the inference neglects a base rate, and so on. These criticisms betray an expectation that our inferences respect what we can call "inferential norms." Proponents of cognitive penetration challenges claim that the subpersonal processing that causes our experiences must also respect such familiar inferential norms, or at least analogous ones. When subpersonal processing violates these norms, the objector says that the violation, contra dogmatism, can prevent an experience from providing prima facie justification for its content. More carefully, what's distinctive about the cognitive penetration challenge is that it raises the question of whether (1) cognitive influence via an unbroken series of subpersonal processing can prevent an experience from providing prima facie justification for its content[12] when (2) some part of the processing violates an inferential(-like) norm. Dogmatism says that it can't; our intuitions say that it can.

You won't need a deep understanding of cognitive penetration's distinctive issue. Feel free to skip to the next section if you already get the gist of it. In the remainder of the present section, I provide some clarification and defense.

The Distinctive Issue Clarified

If A cognitively penetrates B, then A must cause B via an unbroken series of processing. Macpherson's (2012, p. 26) migraine case shows the need for such conditions. Suppose that Fiona believes that she has a test tomorrow. This belief causes her to be anxious, the anxiety results in certain (mere) chemical changes, and the chemical changes trigger spotty vision of the sort associated with migraines. The test belief influences the character and content of Fiona's visual experience, but at least one link in the causal chain (the mere chemical changes) doesn't count as processing.[13] Consequently, no one regards this case as an example of cognitive penetration. And even if we stipulate that

[12] The strongest and simplest cognitive penetration challenges focus on cases in which the experience is cognitively penetrated (i.e., when a cognitive state influences the experience via an unbroken series of subpersonal processing). In principle (see note 15), one could instead argue that the justificatory power of an experience can be affected when some other state is cognitively penetrated.

[13] Heaven help me if pressed for what it takes for something to count as the relevant sort of processing. Nobody seems to have a good answer to that question.

Fiona's test belief is entirely irrational, no one alleges that the test belief's influence is one that prevents the resulting experience from providing prima facie justification for its content. The cognitive penetration challenge, then, depends on the relevant causal influence involving an unbroken chain of processing between the penetrating state and the penetrated experience.

The relevant processing must also be subpersonal. A human person has many proper parts, including fingers and toes. *Subpersonal* properties and processes are properties and processes of a person's parts. Sometimes a subpersonal property/process suffices for a person-level property/process. If Bill's head is bald, then Bill is bald. Yet some properties and processes are merely subpersonal. My parts, in other words, have some properties that I don't have, and they do some things that I don't do. For example, one of my parts is an odd-looking nose. I may be odd-looking, but at least I'm not a nose. Arguably, much of perceptual and cognitive processing is merely subpersonal in this sense. Some part of my brain may use a certain process to fill in the optic disk gap, but I arguably do not perform this process. On the other hand, some properties and processes may be merely person-level. Perhaps whenever I desire coffee, there is no proper part of me that also desires coffee.

My suggestion is that the cognitive penetration challenge applies when there is an unbroken series of subpersonal processing from a cognitive state to, for example, my experience. This is narrow insofar as merely person-level processing is excluded. It is not so narrow, however, that it excludes subpersonal processing which also counts as person-level processing/inference. McGrath (2013, especially section 5) insists that for penetration to affect the justificatory power of an experience, the transition from the cognitive state to the experience must be person-level. But he doesn't deny that the "inferential norm violation" occurs also at the subpersonal level. While I've construed the cognitive penetration challenge narrowly, it still is broad enough to capture all existing cognitive penetration challenges to dogmatism.

The Bad Search Challenge

Overt attention raises a familiar challenge to the idea that undefeated evidence suffices for justification. The basic idea is that undefeated evidence can fail to provide justification when that body of evidence is the result of a bad search for evidence. The bad search can prevent what would otherwise

be evidence that justifies believing proposition P from, in fact, justifying P. Standard examples of bad searches involve bad distributions of overt attention (e.g., focusing on one part of the minority candidate's job application—the part that contains the applicant's weakest credentials—when one should have turned the page and considered a different part of the application). These bad search objections can directly challenge dogmatism. A person's irrational belief that Black people are more violent may affect the way they overtly attend to a given stimulus, making it more likely that their experience (mis)represents the Black person as carrying a gun. Some will argue that the biased distribution of attention prevents an experience that represents the Black person as having a gun from providing prima facie justification for the claim that the Black person has a gun.

It's worth stressing that a subject's merely being biased toward (not) believing P is not sufficient for a search for evidence to be bad. Nor is it sufficient to prevent what would otherwise be justifying evidence for believing P from, in fact, justifying my belief in P. I'm biased toward believing positive things about my children. It simply does not follow that every positive thing I believe about my children is the result of such a bias. I might know that my son just scored the winning goal despite having some tendency to believe positive things about my children even when they aren't true.

Bias threatens (prima facie) justification only when it has some, perhaps indirect, influence on what we believe. In this section, we are focused on cases in which the bias affects what one experiences (and so what one believes) by affecting how one acts, in particular how one searches for evidence. We should also assume in these cases that the subject is reasonably unaware that their experience is the result of a biased search. Otherwise, we can explain the lack of justification by appealing to a defeater—awareness of the biased search—rather than the biased search itself.

The *bad search challenge* to dogmatism claims that when a perceptual experience is the result of a bad search (bad distribution of overt attention), contra dogmatism, the experience may fail to provide prima facie justification for its content. I contend that the challenge's distinctive issue concerns the relation between the practical (what to do/what action to perform) and the epistemic (what to believe).

Where to search for evidence, how long to search, and the manner in which to search are subject to practical considerations, such as the importance of finding out the truth, the costs of further searching, legal and moral constraints on one's search, etc. The example concerning the evaluation of

job applications raises many of these practical considerations. Epistemic considerations also matter for searches.[14] Whether I ought to continue searching for evidence is affected by how much justification I have that further searching may turn up evidence against my current belief that P (see Siegel, 2017a, p. 167). But epistemic considerations cannot make a search bad by themselves. If it's a trivial matter whether P is true and my kid needs to be rushed to the hospital, then it would be insane to continue searching for evidence concerning P, given some chance that further searching would yield evidence that P is false.

If a search for evidence isn't bad, it's hard to see why it would prevent otherwise perfectly good evidence from justifying what it would otherwise justify. It's hard to see why my failure to further scrutinize my belief that P should have any bearing on whether the belief is justified or reasonable, when that failure is due to me rushing my kids to the hospital. If anything, it is the quality of a search that is relevant to whether the resulting experience justifies what it would ordinarily justify.

And if the quality of a search is relevant to whether the resulting experience justifies what it would ordinarily justify, then practical considerations matter epistemically.[15] The quality of a search is always due to the interaction of both practical and epistemic factors. We can always change the quality of the search just by changing the stakes. If my kids are fine and P is all-important, then it would be bad to not search for more evidence concerning whether P is true.

The bad search challenge assumes that the quality of a search partly determines whether the resulting experience provides prima facie justification for its content. Practical considerations partly determine the quality of the search. Thus, we reach the distinctive issue raised by the challenge, namely that practical considerations partly determine whether an experience provides prima facie justification for believing its content. Bad search challenges, in other words, appeal to some version of what has been called

[14] I don't know that anyone has offered a satisfying account of what distinguishes epistemic and practical considerations. It's best to stick with examples of each and hope that you have at least some vague idea of what the distinction amounts to.

[15] https://link.springer.com/article/10.1007/s11098-020-01435-w (2017b, Part III; cf., 2017a, Chapter 9) may insist that biased searches can have a purely epistemic impact, and, contrary to my diagnosis, the bad search challenge need not concern the interaction of epistemic and practical factors. Yet her explanation of how biased searches have purely epistemic relevance explicitly assumes that cognitive penetration can affect whether it is rational to rely on feelings of trust in a search process. Thus, such an approach is no good if we are trying to keep the bad search challenge independent of the cognitive penetration challenge.

pragmatic encroachment: Very crudely put, what you ought to believe is partially determined by practical considerations, such as the importance of finding out the truth on the matter in question.[16,17] Dogmatism is incompatible with this sort of pragmatic encroachment because it holds that experiences provide prima facie justification no matter what practical considerations are at play.

We've been trying to understand the existing challenges to dogmatism so that we can better understand whether covert selection offers a distinctive challenge to dogmatism, one that is importantly different from existing challenges. To avoid collapsing into the cognitive penetration challenge, a covert selection challenge must not appeal to subpersonal violations of inferential(-like) norms. To avoid collapsing into the bad search challenge, it must not appeal to pragmatic encroachment. In the next section, I argue that covert selection challenges don't directly challenge dogmatism's truth without collapsing into one of these other two challenges. In the section following that ("An Indirect Challenge?"), I consider whether covert selection can indirectly challenge dogmatism by challenging an explanatory ambition of typical dogmatists. While I concede that it can challenge the relevant ambition, the challenge is not new. Together, these sections argue that covert selection raises no new (direct or indirect) challenge for dogmatism.

No New Direct Challenge

I'll now argue that covert selection doesn't pose a direct challenge to dogmatism without collapsing into the cognitive penetration or bad search

[16] Feldman agrees that bad search objections raise the issue of pragmatic encroachment, though not in those terms (Conee & Feldman, 2004, Chapter 9, pp. 235–236; cf. Chapter 4, pp. 89–90; Chapter 7, p. 189; Feldman, 2008, p. 347; also see Conee and Feldman 2011, p. 313). He also claims, incorrectly in my view, that the distinction between synchronic and diachronic justification is important in this context (e.g., Conee & Feldman, 2004, Chapter 7, pp. 188–189, as well as Chapter 9, p. 235).

[17] One might wonder whether virtue epistemology can underwrite a bad search challenge without appealing to pragmatic encroachment. I don't have space for a full reply, but here are two things to think about. First, keep in mind that the intellectual character which most directly results in believing P is assumed to be virtuous (we assume the subject has evidence that would justify believing P, were it not acquired in a bad search). The proponent of a virtue-driven bad search challenge must explain why the intellectual character involved in selecting a given past action (how and whether to search for [further] evidence) is relevant to whether one has justification to believe P. My contention is that any such explanation will be committed to pragmatic encroachment. Second, for homework, you can consider Baehr's (2011) virtue-driven bad search challenge and Baril's (2013, section 3.3) explanation of how Baehr's virtue epistemology is committed to pragmatic encroachment.

challenge. Which existing challenge it collapses into depends on which metaphysics of covert selection we assume. Existing accounts of covert selection's metaphysics make covert selection analogous to (1) cognitive penetration and/or inference or (2) overt selection and/or action. In the first subsection, I argue that any challenge to dogmatism posed by the former will collapse into the cognitive penetration challenge. In the second subsection, I argue that any challenge to dogmatism posed by the latter will collapse into the bad search challenge.

Covert Selection as Inference-Like

If covert selection raises a distinct challenge to dogmatism, covert selection would need to be a different kind of thing than cognitive penetration. Insofar as they are the same kind of processing, they are subject to the same epistemic norms. Imagine someone saying, "Covert selection just is cognitive penetration, but we should treat them differently when we do epistemology." This different treatment would seem arbitrary or nonsensical. Mole provides a real-world example of someone who treats covert selection and penetration as having a unified metaphysics and epistemology (2015, pp. 225, 236). He is explicit that covert selection is a "variety of cognitive penetration" (p. 236), and he implies that the epistemologies of cognitive penetration and covert selection are more or less the same (see especially p. 225).

Of course, different kinds of cognitive penetration may be subject to different epistemic norms. The cognitive penetration challenge specifically requires that the problem for dogmatism be caused by subpersonal violations of inference-like norms. Perhaps some kinds of cognitive penetration (those that count as covert selection) are problematic in a way that is not explicable by appealing to anything like the epistemic norms we apply to inferences. Will this suggestion lead to a distinctive challenge from covert selection?

We need to be careful here, lest we trivialize what it means to refer to something as a *covert selection challenge* or a challenge *from covert selection*. Consider the bifold suggestion that (1) the causal history of our experience needs to be reliable in a way that our inferences don't need to be reliable[18] and (2) when covert selection causes an experience to be unreliable in the

[18] As epistemologists use the term *reliable*, it refers to something's tendency to yield true rather than false representations. Part (1) of the suggestion might be cashed out by saying that experiences

relevant way, it prevents the experience from providing prima facie justification to believe its content. At first glance, this suggestion may seem to raise a challenge to dogmatism importantly different than the narrowly construed cognitive penetration challenge. The problem is that covert selection's role in the suggestion is too trivial. The bifold suggestion allows anything that lowers reliability to prevent the experience from providing prima facie justification for its content. A tumor that distorted the retinal signals or an unreliable kind of cognitive penetration would have the same epistemic significance as biased covert selection. If the bifold suggestion raises a problem for dogmatism at all, it does so independently of covert selection. When we are considering whether covert selection raises a distinctive challenge to dogmatism, we are considering whether there are any objections to dogmatism in which covert selection plays the "lead" or the "starring" role in the objection.

Consider the cognitive penetration challenge, as I've characterized it. Cognitive penetration isn't a mere adornment to the objection. Cognitive penetration is a type of causation often thought to be inference-like (see, e.g., Gross, 2017; Pylyshyn, 1999), and it is an interesting question whether such inference-like causal processes can prevent an experience from having justificatory power when those processes violate epistemic norms of inference (or something analogous to those norms). Such objections really are objections from cognitive penetration. Cognitive penetration plays the starring role.

If covert selection raises a challenge to dogmatism at all—if there is to be a challenge from covert selection—then it must play a starring role in some objection to dogmatism. If the challenge is to be distinct from the cognitive penetration challenge, then covert selection's role in the objection must work without appealing to violations of inference-like rules. But once we assume that covert selection is a special type of cognitive penetration, we need to find an objection to dogmatism according to which that special type of cognitive penetration (i.e., the type that counts as covert selection) plays a starring role different than the one ordinarily attributed to cognitive penetration. It's hard to see what that role could be, and no one has offered any suggestions. In the absence of epistemic innovation, we can tentatively conclude as follows: If covert selection is to raise a challenge to dogmatism distinct from the cognitive penetration challenge, then covert selection must be a different kind of thing than cognitive penetration.

need to be caused by a process that is unconditionally reliable, and inferences need only be conditionally reliable (i.e., unconditionally reliable on the condition that the inputs of the inferences are true).

Siegel (2017a, 2017b) holds that covert selection and cognitive penetration are strictly distinct; however, she assumes that the problematic forms of covert selection either involve problematic forms of cognitive penetration or else involve causal histories that are analogous to problematic forms of cognitive penetration. It is no surprise, then, that she traces the problematic forms of covert selection to violations of inferential norms. The more similar we make covert selection and cognitive penetration, the more likely it is that any covert selection challenge collapses into the cognitive penetration challenge.

Covert Selection as Action-Like

Allport's account of attention might be taken as a polar opposite of Mole's view. Recall that, on Mole's view, covert selection is a type of cognitive penetration. Indeed, covert and overt selection are so disunified that they aren't even analogues of each other (Mole, 2015, p. 222). In contrast, Allport treats covert and overt attention as deeply unified. He holds that attention of either sort is a phenomenon at the level of the whole person (see, e.g., Allport, 2011, pp. 25–26, 49–51).[19] It is a mistake to treat any subpersonal process, such as feature binding, as (covert) attention. It is persons who attend, and attention is a relation between persons and that to which they attend.

Once covert attention is thought of as a person-level relation, deeply unified with overt attention, covert selection becomes very different than cognitive penetration. Thus, you might expect a view like Allport's to support a challenge to dogmatism that is importantly different than the one posed by cognitive penetration. And you'd be right. But now the challenge posed by covert selection just is the challenge posed by overt selection, namely the bad search challenge.

The bad search challenge holds, contra dogmatism, that when my bad search for evidence leads to an experience, the experience may fail to provide prima facie justification for its content. The distinctive issue raised by such objections is pragmatic encroachment, the idea that practical factors partly determine epistemic justification. The previous section explained why bad search objections are committed to pragmatic encroachment. But given

[19] Allport and I both allow that organisms can attend even if they aren't persons. I'm just focusing on people here.

Allport's account of covert attention as a person-level relation, that explanation can be extended to show that any covert selection challenge is likewise committed to pragmatic encroachment.

You may doubt that the extension holds if you fail to notice that some relations are actions. If I (intentionally) hug you or kick you, the relation of my hugging you or my kicking you is the action. Hugging and kicking are relations governed by the same norms that govern actions. Allportian covert and overt attention would likewise be governed by the same norms that govern actions. Just as moral or prudential norms might make it (in)appropriate to hug the person next to you, moral or prudential norms might make it (in) appropriate to attend (overtly or covertly) to a certain characteristic of the person next to you. Consequently, practical considerations partly determine whether my distribution of attention is good or bad.

The collapse is caused not by Allport's claim that covert attention is person-level but by treating covert and overt attention analogously. Consider another common way to think about covert selection: It is subpersonal, distinct from cognitive penetration, and it selects what undergoes further processing in an "action-like" way, analogous to overt, bodily action (Gross, 2017, p. 7; cf. Mole, 2015, section 3). The more action-like covert selection is, the more what's selected should be sensitive to practical considerations. For example, suppose a subpersonal mechanism must "decide" whether to submit a certain input to further processing. The quality of this "decision" is not reducible solely to epistemic considerations, such as whether further processing of a given stimulus will increase or decrease reliability. Suppose that further processing would lower the reliability of the resulting experience somewhat by making it more likely that the experience represents a snake when one isn't there. The loss of reliability might very well be worth it if the further processing also would lower the probability that the perceptual experience will fail to represent a dangerous snake when one is there. When covert selection is tightly connected to person-level relations (Allport) or what is submitted for further processing (Gross), practical considerations matter, such as the relative importance of avoiding false positives and false negatives.

Applied to the cases at hand, the argument of the previous section goes as follows. Person-level distributions of attention or subpersonal "choices" of what to submit for further processing prevent a resulting experience from prima facie justifying its content only if the distributions/choices are bad. But the badness of a distribution/choice is not a purely epistemic matter;

practical considerations also matter. Thus, if the badness of a distribution of covert attention/subpersonal "choice" matter epistemically, then practical considerations matter epistemically. In other words, accounts that make covert attention analogous to action challenge dogmatism only by invoking pragmatic encroachment, only by collapsing into the bad search challenge.

What we've seen is that the more similar we make covert selection and cognitive penetration, the more likely it is that any covert selection challenge collapses into the cognitive penetration challenge. The more similar we make covert and overt selection, the more likely it is that any covert selection challenge collapses into the bad search challenge. There is, perhaps, some room for metaphysical innovation that makes covert selection neither inference-like nor action-like. But absent such innovation, we can conclude that covert selection doesn't pose a new challenge to dogmatism's truth.

An Indirect Challenge?

We've seen that covert selection can't provide a direct challenge to dogmatism (i.e., show that dogmatism is false) without collapsing into an existing challenge for dogmatism. In this section, we consider whether covert selection can indirectly challenge dogmatism by, for example, challenging an explanatory ambition held by typical dogmatists. I'll argue that it can, but, yet again, the challenge is nothing new.

The alleged epistemic relevance depends on a controversial account of evidence possession, one which allows past experiences to be relevant counterevidence now (even if you don't remember having them). Suppose I will win a prize if all the squares in a display are red. At time t_0, I begin a visual scan of the display and several seconds later, at t_2, I have a perceptual experience that R (i.e., that all the squares are red). So far, so good; but there's a twist. At some intermediate point in the scanning, I had a perceptual experience of a blue square; but my desire to win the prize prevents that experience from making a difference to subsequent perceptual or cognitive processing, so I do not even remember having the blue square experience. To use Siegel's apt phrase, the desire "anti-selects the experience for uptake" (e.g., 2013). So I now have an experience that represents all squares as red even though a few seconds ago my experience represented a blue square. Siegel argues that I now lack justification to believe R because my past, unremembered experience of the

blue square continues to count as relevant counterevidence for the claim that all the squares are red (2017b, p. 429).[20]

Siegel's suggestion doesn't provide a direct challenge to dogmatism's truth. Dogmatism claims that my experience that R provides justification to believe R if there are no defeaters, no relevant counterevidence. As the case was described, however, my past experience is relevant counterevidence, and thus dogmatism takes no stand on whether I now have justification to believe R. If Siegel's suggestion poses a challenge to dogmatism, it will be in a more indirect way.

Strictly speaking, dogmatism says only that one's (current) perceptual experience that P is sufficient for justification in the absence of defeaters; it comments on neither whether perceptual experiences are necessary for justification nor whether past perceptual experiences can count as current (counter)evidence. Nonetheless, dogmatists often have explanatory ambitions that go beyond the sufficient condition espoused in their dogmatism. For example, some of them aim to explain all justified belief by ultimately appealing to some experience or another (where *experience* is understood broadly enough to include intuitions and apparent memories).[21] If we follow Siegel and hold that past experiences can provide a defeater or relevant counterevidence for current perceptual justification, then one might worry that we are giving up on the spirit of dogmatism.[22]

As you consider the force of this objection, remember what I'm up to in this chapter. I'm arguing that covert attention raises no new challenge for the dogmatist that the dogmatist does not already face. If dogmatists tend to have certain explanatory ambitions, then the objector is correct that

[20] Siegel suggests that, even if the past perceptual representation of the blue square were preconscious, the past preconscious blue square representation could still count as counterevidence. Perhaps, but such a claim depends on the thesis that preconscious states can count as (counter)evidence. That thesis, by itself, is in tension with the explanatory ambition of typical dogmatists; and to the limited extent that Siegel defends the thesis, her defense doesn't appeal to attention. Consider a modification of Siegel and Silins' distracted driver case (2014, 159). In this modified case, the subject successfully drives toward her intended destination, stops and turns when necessary, all the while lacking any conscious experience concerning the road at all. Perhaps the driver nonetheless forms perceptual beliefs about stoplights, curves in the road, etc. Siegel assumes that these beliefs are justified. If she is right, then perhaps a natural conclusion to draw is that some unconscious representation is justifying the belief. This would challenge the relevant explanatory ambition, but appeals to covert selection are not needed to pose this challenge.

[21] This ambition is especially held by those dogmatists who endorse phenomenal conservatism, such as Huemer (2001, 2013a) and Tucker (2010, p. 542, n3; 2011). This ambition is not often stated in print, but I can speak for myself as a proponent of dogmatism that I do hope that experience ultimately accounts for all justification.

[22] Thanks to Hilary Kornblith and Jessie Munton for helping me see the importance of considering this objection.

problems for those explanatory ambitions are problems for at least those specific dogmatists. Yet these problems are nothing new: The relevant explanatory ambitions already face the issues raised by Siegel's first suggestion. In fact, the dogmatist already faces this challenge—call it the *past experiences matter challenge*—in two distinct ways.

Dogmatists with the relevant explanatory ambitions allow that apparent memories/memorial experiences (one's having a conscious episode of seeming to remember such and such) provide prima facie justification to believe their contents. Perceptual dogmatists are, in other words, often memorial dogmatists.[23] Yet certain cases provide a well-known challenge to memorial dogmatism.

Suppose that a few days ago I read that (J) a protester, while waving a sign, matched Congressman Paul Ryan's jogging pace for 9 miles. I believe what I read despite knowing that I'm reading *The Onion* and am well aware of the venue's satirical nature. My belief that J is irrational at the time it was formed because my awareness of the belief's source provided a defeater to the justification I would ordinarily have on the basis of testimony. Yet suppose now that I seem to remember J's being true and I believe J on this basis. The catch is that I no longer recall—I have completely forgotten—the source of my belief in J. I now have no representational state that would plausibly count as counterevidence to my belief that J, as I did when I first formed the belief. Nonetheless, many people have the intuition that my current belief in J is not justified, despite my lacking any current representational state that could plausibly count as a defeater.

The Paul Ryan case and cases like it put pressure on the memorial dogmatist to find a way for past defeaters to continue to have influence on my current justification while remaining faithful to their explanatory ambitions. Note that the explanatory ambition I mentioned—to ultimately account for all justification by appealing to experience—is compatible with past experiences making some sort of difference to current degrees of justification (or lack thereof). One way to grant past experience current epistemic relevance is to allow past experiences to count as current counterevidence, as Siegel suggested; but there may be other ways of doing so that are equally congruent with both our intuitions about which beliefs are justified and the explanatory ambitions of typical dogmatists. What the red square case

[23] Even perceptual dogmatists who apparently lack the relevant explanatory ambition, such as Pollock (1986, pp. 44–52) and possibly Audi (2013, pp. 188–189), also endorse memorial dogmatism.

and the Paul Ryan case show, if anything, is just that past experience can somehow negatively affect current justification. Further argument is needed to determine that the most plausible model for this negative effect (given the explanatory ambitions of typical dogmatists) is by allowing past experiences to play the role of current counterevidence.[24]

I've said that the challenge that Siegel raises to dogmatism's explanatory ambitions is already raised in two distinct ways. First, as we just saw, certain cases put pressure on the dogmatist to allow past experience to negatively affect a person's degree of justification for believing a proposition. Second, dogmatists are under intuitive pressure to allow past experience to positively affect a person's degree of justification for believing a proposition. (While the most salient feature of Siegel's purported moral concerns the negative effect, namely that past experiences can count as current counterevidence to a proposition, this negative effect may depend on the positive effect. For example, one way to have a defeater for believing P is to have evidence for ~P.)

Right now I have justification to believe a great many empirical propositions[25]: propositions about where I live, where my siblings live, how tall my kids are, how much of my office space is devoted to coffee paraphernalia, what color my kitchen is, etc. The problem for the dogmatist is posed by the scope of my current justified empirical beliefs when compared with the scope of my current experience. My current perceptual experience is fairly limited in what it represents—basically just the look of my desk and computer, the lingering smell of coffee and chalk, and the sound of the bad music that I tend to like. It is not currently representing anything about the size of my children, about my siblings, or the color of my kitchen. And before I started thinking about my children, siblings, or kitchen, I had no relevant memorial experiences that provided me with justification to believe any claims about them. Thus, how can the dogmatist account for the full scope of justified empirical belief given the limited perceptual and memorial experiences we have at any given moment?

The dogmatist, then, is under intuitive pressure to allow past perceptual experiences to somehow make a positive difference to perceptual justification now. One way is to allow past experiences to provide evidence now for

[24] See Huemer (1999) for one dogmatist's attempt to deal with these issues in a way that allows past experiences to affect current justification without past experience playing the role of current defeater.

[25] Most propositions about the past, present, and future, including scientific ones, are empirical in the relevant sense. Logical propositions (modus ponens is valid), mathematical propositions (2 + 2 = 4), and moral judgments (it is morally wrong to torture for fun) are non-empirical.

the relevant claims, but we shouldn't assume that this is the only way for the dogmatist to account for the scope of justified empirical beliefs that is congruent with their explanatory ambitions.

Suppose that Seigel's red square case shows us that past experience can count as current counterevidence. I've conceded that, although this result poses no direct challenge to dogmatism, it does challenge an explanatory ambition of typical dogmatists. Yet this challenge is not new: It's one that the dogmatist already faces for independent reasons having to do with the tricky interaction of perceptual and memorial justification.

I've argued that dogmatism already faces direct challenges that appeal to (1) subpersonal violations of inferential(-like) norms and (2) pragmatic encroachment. The explanatory ambitions of dogmatism already face challenges that appeal to (3) the tricky interaction of perceptual and memorial justification. *The take-home message of the chapter is this: There is no way for covert selection to pose a problem for dogmatism or the explanatory ambition of typical dogmatists without appealing to at least one of (1)–(3), and thus there is no way for covert selection to raise a new challenge for dogmatism or the relevant explanatory ambition.*

Conclusion

Dogmatism is the claim that, necessarily, perceptual experiences provide one with prima facie justification to believe their contents (equivalently: necessarily, perceptual experiences provide one with justification to believe their contents in the absence of defeaters). We've seen that dogmatism faces a number of different challenges. I'll forgive the psychologists if they conclude that dogmatism faces too many challenges to be taken seriously; however, in philosophy, all views face many challenges. The goal is to figure out which view is least bad, and dogmatism is in the running for the least bad view of perceptual justification.

The goal was to determine whether covert selection poses a new challenge to dogmatism or one of its explanatory ambitions. I've argued that it does not. Covert selection may raise a direct challenge to dogmatism's truth. Which direct challenge it raises depends on what precisely covert selection is. If covert selection is analogous to cognitive penetration, then it raises the cognitive penetration challenge. If it is analogous to overt attention, then it raises the bad search challenge. Either way, the challenge to dogmatism is old news.

I've also conceded that covert selection may raise an indirect challenge to dogmatism. Dogmatists who aim to account for all justification by appealing to experience may be pressured into allowing past experiences to somehow make a difference to what one is currently justified in believing. But that pressure exists because of tricky issues concerning the interaction of perceptual and memorial justification, and we do not need covert selection to raise this challenge to the explanatory ambition of typical dogmatists.

Dogmatists, like me, have their work cut out for them. To successfully defend their view, they must address a number of distinct challenges. The burgeoning philosophical work on the metaphysics and epistemology of attention makes one wonder whether covert selection could lead to a further challenge for the dogmatist. I have argued that it does not. Phew. That's one less thing I need to be worried about!

Acknowledgments

Thanks to Nathan Ballantyne, David Dunning, Joshua Gert, as well as the audiences at the NYC Epistemology and Psychology Conference and the 6th Annual BELUX Conference, for very helpful comments.

References

Allport, A. (2011). Attention and integration. In C. Mole, D. Smithies, & W. Wu (Eds.), *Attention: Philosophical and psychological essays* (pp. 24–59). Oxford University Press.

Audi, R. (2013). Doxastic innocence: Phenomenal conservatism and grounds of justification. In C. Tucker (Ed.), *Seemings and justification: New essays on dogmatism and phenomenal conservatism* (pp. 181–201). Oxford University Press.

Baehr, J. (2011). Evidentialism, vice, and virtue. In T. Dougherty (Ed.), *Evidentialism and its discontents* (pp. 88–102). Oxford University Press.

Baril, A. (2013). Pragmatic encroachment in accounts of epistemic excellence. *Synthese*, *190*(17), 3929–3952.

Conee, E., & Feldman, R. (2004). *Evidentialism: Essays in epistemology*. Oxford University Press.

Conee, E., & Feldman, R. (2011). Replies. In T. Dougherty (Ed.), *Evidentialism and its discontents* (pp. 283–323). Oxford University Press.

Feldman, R. (2008). Modest deontologism in epistemology. *Synthese*, *161*(3), 339–355.

Gross, S. (2017). Cognitive penetration and attention. *Frontiers in Psychology*, *8*, 1–12.

Huemer, M. (1999). The problem of memory knowledge. *Pacific Philosophical Quarterly*, *80*(4), 346–357.

Huemer, M. (2001). *Skepticism and the veil of perception*. Rowman & Littlefield.

Huemer, M. (2013a). Phenomenal conservatism über alles. In C. Tucker (Ed.), *Seemings and justification: New essays on dogmatism and phenomenal conservatism* (pp. 328–350). Oxford University Press.

Huemer, M. (2013b). Epistemological asymmetries between belief and experience. *Philosophical Studies*, *162*(3), 741–748.

Kelly, T. (2014). Evidence. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2016 ed.). Stanford University. https://plato.stanford.edu/entries/evidence/

Lyons, J. C. (2009). *Perception and basic beliefs: Zombies, modules, and the problem of the external world*. Oxford University Press.

MacPherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, *84*(1), 24–62.

Markie, P. (2005). The mystery of direct perceptual justification. *Philosophical Studies*, *126*(3), 347–373.

Markie, P. (2013). Searching for true dogmatism. In C. Tucker (Ed.), *Seemings and justification: New essays on dogmatism and phenomenal conservatism* (pp. 248–269). Oxford University Press.

McGrath, M. (2013). Phenomenal conservatism and cognitive penetration: The "bad basis" counterexamples. In C. Tucker (Ed.), *Seemings and justification: New essays on dogmatism and phenomenal conservatism* (pp. 225–247). Oxford University Press.

Mole, C. (2015). Attention and cognitive penetration. In J. Zeimbekis & A. Raftopoulos (Eds.), *The cognitive penetrability of perception: New philosophical perspectives* (pp. 218–237). Oxford University Press.

Pollock, J. L. (1986). *Contemporary theories of knowledge*. Rowman & Littlefield.

Pryor, J. (2000). The skeptic and the dogmatist. *Noûs*, 34(4), 517–549.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for the impenetrability of visual perception. *Behavioral and Brain Sciences*, 22(3), 366–423.

Siegel, S. (2012). Cognitive penetrability and perceptual justification. *Noûs*, 46(2), 201–222.

Siegel, S. (2013). Can selection effects on experience influence its rational role? In T. Gendler (Ed.), *Oxford studies in epistemology* (Vol. 4, pp. 240–270). Oxford University Press.

Siegel, S. (2017a). *The rationality of perception*. Oxford University Press.

Siegel, S. (2017b). How is wishful seeing like wishful thinking? *Philosophy and Phenomenological Research*, *95*(2), 408–435.

Siegel, S., & Silins, N. (2014). Consciousness, attention, and justification. In D. Dodd & E. Zardini (Eds.), *Skepticism and perceptual justification* (pp. 149–172). Oxford University Press.

Skene, M. (2013). Seemings and the possibility of epistemic justification. *Philosophical Studies*, *163*(2), 539–559.

Tucker, C. (2010). Why open-minded people should endorse dogmatism. *Philosophical Perspectives*, *24*(1), 529–545.

Tucker, C. (2011). Phenomenal conservatism and evidentialism in religious epistemology. In R. VanArragon & K. James Clark (Eds.), *Evidence and religious belief* (pp. 52–73). Oxford University Press.

Tucker, C. (2013). Seemings and justification: An introduction. In C. Tucker (Ed.), *Seemings and justification: New essays on dogmatism and phenomenal conservatism* (pp. 1–29). Oxford University Press.

Tucker, C. (2014). If dogmatists have a problem with cognitive penetration, you do too. *Dialectica*, *68*(1), 35–62.

Tucker, C. (in press). Experience as evidence. In M. Lasonen-Aarnio & C. Littlejohn (Eds.), *The Routledge handbook of the philosophy of evidence*. Routledge.

8

Bias in a Biased System

Visual Perceptual Prejudice

Jessie Munton

The term *bias* is used in a variety of ways. We often use *bias* colloquially to mean a kind of prejudice against a particular social group. This is the sense in which we talk of *implicit bias* grounding discrimination or of a system that is unfairly *biased* against a racial minority, for instance. This sense of bias is inherently freighted with a negative valence: It has problematic ethical upshots, and it is also often assumed to be epistemically flawed, either grounded in inaccurate information or irrational cognitive processes or liable to perpetuate flawed reasoning downstream of it.[1] We can call this kind of bias *prejudicial bias*. Prejudicial bias generally concerns demographic groups: people of a particular race, gender, sexual orientation, or class.

But *bias* also has a thinner, formal sense, meaning any kind of weighting in a testing process that systematically skews the outcome. This is what we have in mind when we talk of a *confirmation bias*, for instance, when a process is skewed toward producing results which confirm a previously endorsed hypothesis. Bias in this second sense may sometimes also be prejudicial against particular demographic groups, but it need not be. And some kinds of action recommended to combat the first, negatively valenced kind of bias may be instances of the second, formal sort of bias, such as deliberately assigning readings by female authors on a particular topic to combat their underrepresentation on philosophy syllabi or practices of affirmative action. We can call this second kind of bias *formal bias*.

Formal bias need not always constitute an epistemic flaw, though it may sometimes do so. Under the right conditions, formal biases maximize the

[1] For a range of philosophical perspectives on the nature of implicit prejudicial bias and the sorts of problems it can give rise to, see Brownstein and Saul (2016).

Jessie Munton, *Bias in a Biased System* In: *Reason, Bias, and Inquiry.* Edited by: Nathan Ballantyne and David Dunning, Oxford University Press. © Oxford University Press 2022. DOI: 10.1093/oso/9780197636916.003.0009

information at the disposal of an organism. It is not, therefore, inherently valenced, either epistemically or morally. In this sense, it is a neutral category.

What is the relationship between prejudicial bias and formal bias? Is prejudicial bias just an instance of formal bias that happens to have a particular demographic orientation, or is it a category that is marked out itself by further structural flaws in reasoning, for instance? This question has important upshots for our understanding of the epistemic status of prejudicial bias: Are the epistemic problems that frequently seem to accompany it merely contingent, or is it in fact distinguished by structural properties which are themselves constitutive of an epistemic flaw?

Standard accounts of epistemic normativity have tended to abstract away from the particular content of the beliefs or arguments under evaluation in favor of focusing on formal requirements they must meet to attain a certain standard, such as knowledge or justification.[2] Evidentialism, for instance, claims that believers must conform their beliefs to the evidence. This requirement applies irrespective of the subject matter of the beliefs in question (Conee & Feldman, 2004). Similarly, reliabilism about knowledge or justification emphasizes the significance of accuracy or the truth ratio of a given means of producing belief (Goldman, 1979; Kornblith, 2002). This is again a formal requirement: It does not pertain to the contents of the beliefs under scrutiny. If prejudicial bias is epistemically flawed as a category, then we would expect it to be distinguished not just by its demographic focus but in addition by certain structural features in virtue of which those flaws arise since standard methods of epistemic evaluation are not sensitive to whether or not the content of belief concerns a demographic group.

In this chapter I pursue the question of whether there is some structural feature that distinguishes prejudicial bias from the broader class of formal bias, specifically in the context of the visual system. Visual perception is a process which demonstrates systematic formal bias. It relies on previously encountered information to guide its accumulation and interpretation of new data. There is also evidence that visual perception can demonstrate a sensitivity to demographic categories including race and gender, in a manner

[2] The exception to this general rule is the literature on pragmatic encroachment, the view that what a subject has practically at stake can "encroach" on epistemic standards applied to them (Fantl & McGrath, 2002; Hawthorne, 2003; Stanley, 2005). Higher stakes make it harder to know a given proposition. Applying this variable standard requires attention to the contents of the belief and details of the believer's situation.

that intuitively resembles a kind of prejudicial bias.[3] Is there a way of distinguishing these instances of prejudicial bias within a formally biased system? I argue that there are cases of prejudicial bias which are indistinguishable in formal terms from the standard operation of the visual system. But I also claim that these cases do manifest distinctive epistemic flaws.

Understanding how we can square this circle requires a reappraisal of our epistemic evaluation of visual perception more generally. To capture the full range of instances of visual prejudicial bias we need to conceptualize visual perception as an active practice that develops over time, that crucially involves not just transitions between pieces of information but decisions about what information to acquire, and that can consequently only be fully epistemically appraised relative to a goal and an environment. The introduction of these additional parameters gives us the room we need to individuate prejudicial bias in a way that reveals it to be constitutively tied to a distinctive kind of epistemic flaw. In doing so, we make some progress on the broader question of how to identify prejudicial bias within a structurally formally biased system.

This chapter proceeds as follows. In the first section I describe how the visual system is both formally biased at a structural level and capable of engaging in a kind of prejudicial bias. I draw on two bodies of empirical work—on face perception and the resolution of perceptual ambiguity—to make this point. In the second section I consider candidate criteria that fail to demarcate prejudicial bias from mere formal bias. In the final section I offer an analysis of why these criteria are bound to fail and propose a different approach to visual perception, one that emphasizes its status as a skill honed through active practice in a given environment, whose evaluation is only possible relative to a set of goals.

Vision as a Biased System

The visual system faces two major challenges. On the one hand, it has too much information: The information available from the environment at any

[3] We tend to think of prejudicial bias as something that manifests in cognition, and that is, if not the upshot of person-level processing, at least something for which an individual may be blamed or held accountable. I am using prejudicial bias in a thinner sense that can apply to the upshot of any process realized by an individual, a subsystem of an individual, a computer algorithm, or some other kind of system.

one moment far surpasses what the brain can process (Summerfield & Egner, 2009). On the other hand, it has too little information: the three-dimensional world is projected onto the two-dimensional retina in a manner that is essentially ambiguous, and the brain must work to reconstruct a representation of the environment from that impoverished data (Scholl, 2005). The solution to both of these problems is similar: The visual system learns from past exposure. Relying on previously experienced environmental regularities allows it to efficiently select relevant information and resolve indeterminacy in retinal data. In some sense, the visual system is thereby systematically biased. Specifically, it is biased toward the interpretation of the novel in line with the familiar.

More particularly, the visual system overcomes the uncertainty inherent in retinal stimulation by relying on a set of priors that guide its interpretation of new data.[4] To take a simple example, suppose you are looking at a line. The retinal space occupied by that line is consistent with it being a number of different lengths, depending on how far it is from you and the angle at which it is positioned. The visual system can resolve that uncertainty by drawing on prior probability distributions over possible values for its length, angle, and distance from you. It can then calculate the most likely value for each of those parameters and in so doing arrive at a more or less determinate representation of that aspect of its environment.[5]

Relying on priors in this way increases the chances of the visual system accurately representing its environment and finding relevant information. It also introduces a kind of minimal confirmation bias since ambiguities in novel stimuli are resolved in line with previous regularities. We do not come at the world with fresh eyes every time. Instead, our perception of the new is colored by our perception of the old.

But note that although this reliance on priors is an instance of a process that is formally biased, it can still be optimally rational. In fact, it can provide a counterweight to prejudicial bias, in so far as proper attention to base rates often forestalls the formation of prejudicially biased belief. Attending to just how unlikely it is that one would be robbed at all, for instance, makes it irrational to have anything but the lowest confidence that an approaching male

[4] This process is formalized in Bayesian models of visual perception, which claim that the visual system integrates old and new information in line with Bayes's theorem. According to Bayes's theorem, the probability of a given hypothesis given a piece of evidence is equal to the probability of the evidence given the hypothesis, multiplied by the probability of the hypothesis: $p(h \mid e) = p(e \mid h)\, p(h)$ (Feldman, 2014; Rescorla, 2015).

[5] For related examples described in considerably more detail, see Kersten and Yuille (2003).

is likely to rob you, despite the higher likelihood that a man rather than a woman would rob you, were you to be robbed.

The visual system demonstrates this kind of formal bias not just in the resolution of uncertainty but in the processes of selecting visual input. Context primes subjects to identify scene-consistent objects—bread in a kitchen, for instance, or a football player on a football field—because those are the contexts in which we have encountered those objects previously (Davenport & Potter, 2004; Palmer, 1975). It can similarly help us predict where to look within those scenes to find relevant stimuli (Wolfe & Horowitz, 2017).

This process of learning to prioritize certain pieces of information gives rise to the development of a kind of perceptual expertise, the development of fine-grained capacities of discrimination and recognition with frequently encountered stimuli (Gauthier et al., 2009). We can recognize instances of this specialization when reflecting on our own visual experiences: The novice gardener has to inspect each plant sprout carefully to distinguish weed from seedling, whereas the expert gardener has no such difficulty; the radiologist now easily parses X-rays into the body parts they represent when at first they seemed a confusing mass of black and white. These effects rely in part on the direction of attention but also on changes to lower-level perceptual processing (Harel et al., 2013): The visual system is capable of changing in response to previous tasks and encounters. This expertise is once again an instance of formal bias: What information is extracted from a particular scene depends on the system's prior encounters.

And yet, this kind of formal bias has the potential to support a form of prejudicial bias, when the visual system draws on stored information not just about low-level features of its environment, such as the likely length and incline of lines, but in addition about higher-level features, such as race or gender. In this section I briefly outline two different bodies of empirical work in visual perception which suggest this is a possibility: work on face perception and work investigating how stereotypes influence the identification of objects under time pressure.

Prejudicial Bias Through Facial Expertise

Our capacity to recognize and read information from faces is a revealing window on the visual system's sensitivity to race. There is evidence that different faces and genders are coded for by different neural populations.

Repeated exposure to faces with artificially distorted features results in a kind of visual desensitization: Normal faces, viewed subsequently, look distorted.[6] This kind of adaptation to Caucasian faces with artificially distorted features does not transfer to Chinese faces and vice versa. Similarly, adaptation effects are specific to male or female faces (Jaquet & Rhodes, 2008; Little et al., 2005). This happens in a manner that suggests a sensitivity not just to low-level physical features but to the social category of the face (Jaquet et al., 2007, 2008).

In fact, there is good evidence that our skills at face perception are systematically arranged along racial lines. We are more accurate at recognizing own-race and dominant-race faces, with both fewer false positives and negatives (Meissner & Brigham, 2001). A White person in a majority White society will generally be better at recognizing other White faces than Black or Asian faces, for instance. A Black person living in a majority Black society is likely to be less good at recognizing White faces than Black faces. Disturbingly, the difference in performance is not limited to recognition: Subjects' identification of emotion is also more accurate for same-race faces (Elfenbeim & Ambady, 2002).

This differentiation emerges in infancy: Kelly et al. (2007) describe a process of "perceptual narrowing" that emerges over the first 9 months of life and involves a loss of capacity to recognize other-race faces. While 3-month-old Caucasian infants could recognize White, Asian, and Black faces after a brief period of exposure to color images of them, 6-month-olds could recognize only White and Asian faces, and 9-month-olds could recognize only Caucasian faces.

What are the epistemic upshots of this pattern of coding facial information?[7] One is that it allows for the extraction of more, and more fine-grained, information from faces. We become expert at the sorts of face we most commonly see. That expertise seems like a positive result, in so far as we get more information than if we did not develop it. But another upshot of it is that there is a discrepancy in our ability to acquire information about people depending on their race. We recognize some people less readily than others and read information from their faces less easily. This discrepancy itself is significant: It leaves us better equipped to interpret and interact with individuals of one race than another. Additional information about Caucasian faces,

[6] This is an adaptation effect, akin to the "waterfall effect": After viewing a waterfall for a period of time, the still ground around one can appear to move upward (Clifford et al., 2007).

[7] For further discussion of the epistemic costs of same-race face effects, see Gendler (2011).

say, comes at the cost of racial neutrality, and neutrality itself may be epistemically valuable in certain situations.

One problem this gives rise to is a kind of snowballing effect of the expertise: Dominant-race individuals extract more information from same-race faces. As a result, they are better positioned to further interact with those individuals and to cultivate relationships with them through improved skills at identifying emotions, for instance. That in turn is likely to further skew their facial perception skills toward reading same-race faces.

Further difficulties are added when we consider that this failure of neutrality does not advertise itself to the subject of the experience. This is especially concerning given the discrepancy in reading emotion from same- and different-race faces. After the same exposure to faces of two different races, they will come away with different quantities of information, facilitating better recall and identification of emotions in one case than the other, while being liable to think of themselves as having equal access to information from both faces. That in turn encourages inaccurate inferences about the resulting discrepancy in the information they have: that other-race faces display less emotion, for instance.

But this structure of expertise also has a degree of ecological validity: Infants' skills at recognizing and reading the faces around them scaffold their developing social skills. A similar specialization is to be found in a range of other contexts as individuals become skilled at reading information from X-rays or livestock. Yet the same-race face effect seems like an instance in which the visual system demonstrates and thereby perpetuates a kind of racial bias. This raises the following question: What distinguishes face recognition as an instance of prejudicial bias, against a backdrop of formally biased task-specific expertise within the visual system more broadly?

Prejudicial Bias Through the Speeded Recognition of Ambiguous Stimuli

A second way in which visual perception could manifest a kind of prejudicial bias against certain demographic categories springs from its reliance on priors as a way of resolving ambiguity in incoming data. Could associations with racial categories, or previous regularities, prime one to identify properties stereotypically associated with race? If that were the case, the visual system could in effect recapitulate prejudicial racial biases, by interpreting

incoming retinal data in line with previously encountered regularities between different races and certain contexts. Suppose, for instance, that you work at an academic institution where only a small percentage of the ladder faculty are Black or Hispanic. When you come to eat in the dining hall, however, those proportions are roughly inverted among the catering and maintenance staff. If your visual system encoded those context-specific regularities, it could interpret novel data in line with them: The ambiguous face at the periphery of your visual field might be resolved as White when you are in a faculty seminar but as non-White when you are in the dining hall. In allowing for that stereotype-consistent resolution, your visual system would in effect be racially biased, eliding information that contradicted its pre-existing racial associations.

The possible impact of these kinds of effect becomes more disturbing if we countenance the possibility that your visual system is sensitive not just to regularities in the real world but to regularities encountered in the media, via photographs or film, through television and the internet. This opens the door to a far wider scope of stereotypical associations, via the regular association of violent crime and young Black men through news reporting.

It is a disturbing thought that the priming of certain concepts, or a certain pattern of past exposure, could determine what we end up seeing as we look at the world around us, in a manner consistent with racist stereotypes. Could such influences make you more likely to identify ambiguous objects as crime-related, for instance? A variety of recent work in vision science purports to demonstrate that racial stereotypes may influence visual perception.

Take a representative study by Correll, Park, Judd, and Wittenbrink (2007), in which subjects played a simple video game. A series of photos, each showing a man in one of various contexts (in front of a car park, on the street, in a park), appeared on the screen before the subject. In the images the man was holding either a gun or a harmless object such as a cell phone. The subjects' task was to indicate "shoot" in response to those individuals who are holding weapons and "not shoot" in response to those holding innocuous objects. Subjects responded under time pressure by pressing a computer key to indicate their choice. Correll and collaborators were interested to see whether the speed and pattern of errors in subjects' responses were sensitive to the race of the target individual featured in the photo. The results revealed a kind of racial bias: In

line with stereotypes associating Black men and crime, subjects were quicker to select the "shoot" response for images of Black men holding guns and were more likely to mistakenly shoot Black men than White men. One important question this gives rise to is whether the subjects just responded in a manner consistent with racial bias or whether their visual experience of the stimulus itself was partly responsible for this pattern of results.

Earlier work by Jennifer Eberhardt et al. (2004) suggested that effects of this kind could be genuinely perceptual: Eberhardt et al. tasked subjects with identifying photographs of objects as quickly as possible. The images were degraded with visual "noise," which gradually cleared, making it progressively easier to identify the objects in question. Eberhardt et al. found that subjects were quicker to identify crime-relevant objects, that is, through a greater quantity of visual noise, when primed with a Black face than with a White face. In this case it seemed that the difference in performance had to be attributed to them seeing the relevant objects sooner, through the noise laid over them.

By contrast, work by Keith Payne suggested that the kind of effect demonstrated by Correll et al. could best be explained not in terms of the impact of racial stereotypes on visual perception directly but merely on the individual's capacity to control their responses. Payne (2001) had initially found that priming individuals with a Black face as opposed to a White face made them more likely to identify an image as a gun than as a tool. But in later work he found that when the time pressure to respond was alleviated, subjects' stereotype-consistent errors evaporated. And subjects could almost always identify when they had made such an error, suggesting that it was a problem at the level of response rather than perception (Payne et al., 2005).

Correll and his colleagues (2015), hoping to resolve these competing interpretations of the influence of racial associations on perceptual experience, performed a "diffusion analysis" on their earlier data. This process uses subjects' accuracy and latency to model their decision-making process. In particular, it aims to pull apart the relative significance of three possible points at which the biased pattern of response could emerge. Firstly, it could be that subjects start from a position that is already biased toward a particular response tendency (i.e., shooting) when confronted with Black targets. Secondly, it could be that they move more quickly from their starting point

to the point at which they have apparently accumulated sufficient evidence to respond on stereotype-congruent trials. Or, thirdly, it could be that the difference in response is accounted for by a speedier action response, after that process of accumulating information has been completed.

Correll et al. (2015) found that the difference was in the rate at which subjects "accumulated evidence," that is, the time it took for them to reach a point at which they were willing to make a decision. For armed targets, participants accumulated evidence more quickly when the target was Black than when the target was White. This difference in the "drift rate" suggests a perceptual element to the effect since visual perception was the key means by which subjects accumulated the evidence in question.[8] Correll et al.'s verdict was as follows:

> Overall, then, the results from diffusion model analysis suggest that participants accumulate evidence more quickly when targets "fit" prevalent stereotypes, but more slowly or gradually when targets violate those stereotypes. This pattern suggests that the targets' race may guide visual interpretation of the object, perhaps by offering supplemental information. (2015, p. 225)

In the same paper, the authors also addressed the question of how stereotypes contributed to the pattern of biased responses by tracking subjects' eye gaze. Since acuity falls off sharply within a few degrees of the fovea, the most light-sensitive part of the retina, where subjects look tells us what information they are prioritizing while undertaking the task. Given that the task was to differentiate the object the target was holding, one would naturally expect that subjects would in all cases look directly at that target object. Surprisingly, Correll et al. found a difference here that depended on the

[8] Interestingly, not all subjects demonstrate racial bias in these tasks. Correll, Park, Judd, Wittenbrink, et al. (2007) compared the responses of police and civilian populations and found that while police officers responded faster for stereotype-congruent trials, they did not manifest racial bias in their ultimate decision to "shoot" or "not shoot." The authors speculate that officers' training impacts on the placement of the ultimate decision criterion; that is, it inhibits the shoot response in ambiguous cases, rather than impacting on the speed with which stereotype-congruent or -incongruent targets are processed. It is worth noting that Plant and Peruche (2005) found that police officers initially did manifest bias in their responses to unarmed Black and unarmed White suspects but that this could be eliminated with task-specific training. The training again seemed to work by inhibiting racial concepts, as revealed by responses on a word-completion task after training on the program.

race of the target. The visual angle (the angle between the subject's visual focal point and the relevant object) was significantly greater for Blacks than Whites, suggesting "that participants shot Blacks with relatively low visual resolution or clarity concerning the object, whereas they achieved much greater visual resolution before shooting an armed White" (2015, p. 227). When the target was Black, subjects were attending to other parts of the image, such as the face, rather than looking directly at the relevant object. Correll et al. offered the following rationale for those results: "[i]f race augments the available visual information on [stereotype-congruent] trials, participants should require less of the available *objective* information. . . . If a gun in the hands of a White man somehow looks less readily like a gun, participants . . . should seek greater clarity through an extended visual search" (2015, p. 225, original italics).

The authors' overall verdict is as follows:

> This is exactly the pattern we would predict if racial stereotypes augment visual processing, leading participants to more quickly interpret ambiguous evidence, such that they reach a stereotypic decision more quickly (as measured by the drift rate index) and so require less fine-grained information (as measured by the visual angle index). (2015, p. 228)

Accepting the authors' interpretation of their work, this looks like a case of prejudicial racial bias: The visual system is precluded from gathering relevant information that would counteract a prejudicial association, by the influence of that very association. In what follows, I will call cases of this sort, in which prejudicial bias arises from a reliance on a prior expectation, "Correll cases." Our condemnation of the structure of influence in these cases is complicated by an appreciation of the systematic reliance of the visual system on stored information, in the manner I have described. We know that, more generally, objects appearing in a consistent background context (a loaf of bread on a kitchen counter) are identified more quickly than when they appear in incongruous contexts (a drum in the same setting) (Davenport & Potter, 2004; Palmer, 1975). In these cases too, the visual system draws on a "stereotype", an association between contexts and objects, just as it does in the Correll cases, in which it relies on an association between race and crime. Is there some unique marker which distinguishes cases of prejudicial bias and explains our sense that they are epistemically problematic?

Candidate Criteria

Irrational Transformations

It can seem obvious that prejudicial bias, even in the perceptual case, involves *bad reasoning* (or its subpersonal, perceptual equivalent): the overweighting of certain past experiences, resulting in inappropriate priors or a failure to appropriately integrate them with retinal data. Susanna Siegel (2013b, 2016) argues that we can rationally appraise the subpersonal transitions involved in the formation of perceptual experiences. Irrational transitions have a negative impact on the capacity of the resulting visual experience to justify belief, just as irrational transitions in the formation of belief diminish its rational power. Instances of prejudicial bias could be distinguished by deviant processes of updating on prior information. According to Siegel's proposal, this would in turn reduce the epistemic power of the resulting experiences, that is, their capacity to justify belief.

This offers us a handle on the Correll cases. There could be rational flaws at many stages in the process which leads up to the relevant experience: The visual system might draw on priors that associate Black men and crime to a degree that is out of all proportion with the evidence encountered. Or it might put undue weight on stereotype-congruent priors, letting them "hack" the process of interpreting incoming data. As a result, the "conclusion" of that process would be disproportionate to the legitimate data available to the system.

This criterion is likely to distinguish a large proportion of instances of prejudicial bias, by assimilating bias to an existing category of epistemic flaw: irrationality. But we might worry that it will not catch every case that intuitively falls within the set of prejudicial bias. For starters, this standard has little say in the case of facial expertise, where the problem can't easily be characterized in terms of transitions between subpersonal states. Moreover, it may not pick out every Correll-style case of visual prejudicial bias. To see this, take the Correll case again, and suppose that a stereotype that does reflect the individual's evidence has been drawn on, proportionally, to inform the relevant experience. Such a case need not fall foul of evidentialist norms: The visual system is responding proportionately to the information at its disposal. In fact, to achieve the kind of intuitive neutrality we instinctively favor, what we want is for the visual system to substantially disregard some of the evidence it does have in favor of evidence it does not have. Equally, to take

another candidate rational norm, prejudicial bias need involve no failure of coherence on the part of the individual or their visual system (BonJour, 1985; Lehrer, 1990). On the contrary, the problem seems to be rather that coherence is playing too large a role in informing the resultant state: Coherence of visual experience with prior expectations itself strikes us as a flaw.

Can the problem be captured instead in terms of reliability? This brings us back to the observation we started with, that patterns of structural bias are an integral element in the normal functioning of the visual system. Given that fact, we are confronted with a nasty instance of the generality problem: How is the process responsible for prejudicial bias to be typed?[9] If we characterize it in formal terms, it is likely to be of the same type as various other processes in visual perception that involve a similar use of priors in the interpretation of incoming information. That leaves us with no distinguishing marker between cases of prejudicial bias and normal functioning. How else can we type the process so as to avoid this result? Must we appeal to content as a way of distinguishing the relevant processes? This lands us back where we started, trying to find a distinctive structural feature, beyond content, that marks out instances of prejudicial bias within the visual system.

Neglect of Available Information and Cutting off Enquiry

One particularly troubling feature of both the same-race face effect and the Correll cases is the way in which the subject has evidence at their disposal to which they fail to give adequate weight. If the subjects in the Correll case paid greater attention to the target object (in preference to the face of the individual holding it), they might gather counterevidence to the stereotype they rely on. Similarly, same-race face expertise results in the systematic neglect of information from other-race faces. These practices of visual search are significantly neglectful of available information. They are akin to someone who claims an interest in education policy but who fails to open the book on their desk on precisely that subject, favoring instead to largely report their pre-existing beliefs. Perhaps prejudicial visual bias involves a distinctive pattern of

[9] The generality problem is the problem of arriving at a principled means of "typing" the processes responsible for belief, in a manner that gives rise to a determinate assessment of their reliability (Conee & Feldman, 1998). The reliability of the relevant process determines the degree of justification the belief enjoys.

selection effect, resulting in the epistemically problematic neglect of available information.

Susanna Siegel (2013a) considers some of the epistemic impacts of selection effects in her discussion of visual experiences that are "anti-selected for uptake": Suppose someone has a visual experience of eggs in the fridge but fails to draw on that information when acting or reasoning—they go on to buy more eggs, for instance. Siegel argues that the neglected experience retains evidential relevance and constitutes a defeater. In line with this, perhaps visual prejudicial bias involves a failure to take up certain elements of an experience for processing. That neglected information serves as an epistemic defeater for the resulting experience.

Siegel's version of this response is couched in terms of the neglect of experiences which are already available to the individual in some form. As it stands, this will leave untouched cases in which the individual has no such experience. This matters: In the problem cases we have considered the relevant information from other-race faces or from the experimental image of the target holding the object need not be processed at any level. It needn't be the case that the subject has a fine-grained experience of an other-race face and then fails to "process" relevant information from it. They may only ever have a coarse-grained experience of it, one that fails to deliver the information required for subtle emotion identification.

So we need to tweak this response to allow that information that never even features in an experience, but that could easily have done so, can serve as an epistemic defeater. What seems culpable in these cases is the fact that subjects don't have such an experience, when they could have done so, had they only directed their attention to the relevant information. Perhaps the flaw then is that they cut off enquiry too soon.

Could prejudicial bias involve a distinctive neglect of available information, prompted by the subject's pre-existing attitude toward the question that the visual search is intended to settle? That could explain in turn the various epistemic flaws prejudicial bias is liable to give rise to: The resulting experiences are less reliable, for instance, because relevant information was neglected.

A standard of this kind would not be an ad hoc development for the visual case. Nathan Ballantyne (2015) argues for a broader norm for belief that allows that our recognition that there is unpossessed evidence against our views can serve as a defeater. Still, this approach fails to fit our purposes in two ways. In the first case, Ballantyne's arguments concern unpossessed evidence

of which a subject is aware. They therefore have evidence of a defeater for the relevant belief. The presence and problems of prejudicial visual bias seem consistent with such evidence not being in any sense available to the subject. Secondly, this account overpredicts. We are almost always aware of unpossessed evidence relevant to a given belief. Ballantyne is open to the possibility that this grounds a kind of broad skepticism about the epistemic status of our beliefs. But as we seek a norm capable of distinguishing cases of prejudicial bias, the global aspect of this skepticism is unhelpful. We want a feature specific to these cases.

This helpfully points our way forward, however: The difficulty posed by these cases is precisely that of pinning down when such neglect of information is illegitimate. We need to ask why we neglect information, and the worry is that the answer will not always negatively impact on the reliability of the visual system or constitute any deviation from the visual system's usual store of processing methods.

Consider cases of dramatic neglect of unexpected stimuli, as occurs with *inattentional blindness*, when observers routinely fail to spot an unexpected stimulus while engaged in a visual task that requires them to extract information of a certain kind from a display. For instance, when counting passes in a basketball-like game, observers miss a man in a gorilla suit walking across the court (Chabris & Simons, 2010) or when counting touches between gray items they fail to notice a bright red cross moving across the field (Ward & Scholl, 2015). In these cases, the visual system faces a trade-off between the information gained by devoting resources to spotting very surprising stimuli and the information won by focusing instead on the task at hand and neglecting irrelevant stimuli. In general, neglect of information is not necessarily irrational when your processing resources are limited. Such neglect cannot, then, serve to distinguish prejudicial from structural bias.

Perhaps prejudicial bias arises when differential quantities of attention are paid to a subset of stimuli, tracking a distinction that is irrelevant to the task at hand. Intuitively the problem with race-based facial expertise, the feature that makes it a potential case of prejudicial bias, is the difference in subjects' response to own-race and other-race faces. Similarly, the discrepant response to White and Black men in the test stimuli is what troubles us about the Correll cases. The subject's task is to determine whether the target is holding a gun, not to determine anything directly related to their race. It is unwarranted, then, to adopt different strategies when confronted with White and Black individuals. Perhaps avoiding prejudicial bias requires us not to attend

to neglected information but the opposite: to shield irrelevant information from influencing the process in question, as in name-blind job applications or gender-blind orchestra auditions.[10]

The difficult question with which this strategy confronts us is, what makes a difference relevant or irrelevant to the task at hand? The explicit description of the Correll task makes no mention of race. But the individual's prior, we can hypothesize, attributes some higher likelihood of manifesting the relevant property, carrying a gun, to individuals of one race than another. Is it always illegitimate to draw on a prior association with a property not explicit in the description of the task? That seems too strong a restriction to place on the use of prior information. If I am engaged in a search for strawberries, I will do well to draw on a prior that strawberries are red, even though my search is not for strawberries under a description of them as red. To prohibit appeal to an association in virtue of its racialized content is to fall back to a characterization of prejudicial bias in contentful rather than formal terms.[11]

We need, then, some further reason to think that this kind of information is not legitimate, that the supplementation with particular information, from stereotypes, for instance, does not give the subject good reason to stop their search. That is the possibility I consider next.

Arational, Emotive Attitudes

Another possible criterion appeals to facts about the kind of state that influences visual or cognitive processing. On this approach, prejudicial bias involves the neglect of information driven not by proportionate priors but by arational affective attitudes, for instance.

Perhaps the problem in the Correll cases lies in the visual system's appeal to states such as stereotypes. I will take a stereotype to be a cluster of information associated with a particular social group, information which can take a variety of forms including propositional beliefs and affective attitudes. Should we assume, for our purposes, that stereotypes have a distinctively negative epistemic valence? The existing literature is split on this question,

[10] I am indebted to David Dunning for this suggestion.

[11] Work in machine learning gives rise to cases that raise similarly difficult questions of when indirect tracking of racial properties via "proxy" properties (properties that correlate with, while not explicitly mentioning race) amounts to racial bias. For an argument that this kind of reliance on proxies could serve as a model of implicit bias in humans, see Johnson (2020).

sometimes treating the category as epistemically neutral and sometimes assuming that stereotypes are, by definition, epistemically flawed.[12] We face a dilemma whichever way we go on this question. If we use *stereotype* to indicate a distinctively inaccurate collection of information, then the problem with its involvement in these cases lies not in its classification as a stereotype but in its inaccuracy. If, on the other hand, we draw the class of stereotypes such that it includes accurate instances, then we are owed an explanation of why such states cannot legitimately serve as stores of prior information. In neither case does the appeal to the notion of a stereotype per se give us additional traction on the problem at hand.

What we have in mind by appealing to stereotypes to mark out cases of prejudicial bias may be not their inaccuracy but the possibility that they involve arational affective attitudes. Affective attitudes may also be in play in the case of face perception. Expertise with own-race faces is partly driven by the close emotional attachments infants and children develop with their caregivers. We can set aside the label of stereotype and look instead just at this possibility: that prejudicial bias is distinguished by the influence of arational affective attitudes on cognitive processing. In virtue of their arationality, such affective attitudes are liable to "hack" the processes by which new and old information are integrated. This could result in beliefs or experiences that fail to be proportionate to the evidence and that are instead skewed to support the existing emotive attitudes of the individual.

Pinning the blame on the drive of an affective attitude requires us to point in turn to a principled reason why affect is an invariably illegitimate influence. Affective states can encode previously encountered information. Consider the fear you feel entering notoriously shark-infested water. That fear is an affective attitude, but it can also be proportionate to the evidence you have of the risk of shark attack.[13] And it can, moreover, skew results in ways that seem epistemically beneficial: You are now more likely to detect sharks because your fear makes you appropriately sensitive to possible stimuli that you would not normally spot, such as fin-like protrusions or gray shapes lurking beneath the surface of the water. There is room for these attitudes to exert a legitimate influence on the direction of attention despite their affective form.

[12] See the work of Lee Jussim and his collaborators for arguments that stereotypes are frequently accurate, in the sense that they are statistically borne out (Jussim et al., 2009; Madon et al., 1998). Others take their inaccuracy to be a definitional feature of them (Blum, 2004; Jost & Banaji, 1994).

[13] See Wells and Matthews (2014) for a full discussion of the role emotion plays in directing perceptual attention. For a discussion of the capacity of emotions to constitute a proportionate or rational response, see Turski (1994).

If the involvement of affective attitudes alone were sufficient for prejudicial bias, the shark case would have to also be counted as an instance.

All of these potential criteria will undoubtedly circumscribe some set of cases, some instances of which will be particularly problematic, embodying prejudicial bias. But none of the criteria seem to catch only classes of prejudicial bias, nor do any of them seem capable of catching all cases of prejudicial bias. For in every case we can imagine an instance of prejudicial bias which does conform to the standard in question but still has prejudicial upshots. And similarly we can imagine innocuous cases of merely formal bias which are caught by the relevant criterion.

A Skill-Based Account of Visual Prejudicial Bias

Why are these epistemic norms failing to capture the phenomenon in question? One reason is that standard forms of epistemic evaluation focus on transitions between pieces of information, or the truth ratio of processes that govern such transitions. In this way, they apply to processes that operate over a fixed body of information. We naturally think of perception as though it were such a process: The immediate environment is fixed, and that determines the information available to the open eye as it views it. But that natural thought is misleading. Given the processing limitations on visual perception, it has to be selective even within a fixed environment. As a result, a significant part of the perceptual process involves a series of subpersonal "decisions" about which available stimuli to focus on. What we need in order to capture the cases of visual prejudicial bias that have been described is an evaluation of the processes by which certain pieces of information, and certain interpretations of them, are prioritized over others.[14]

But it is hard to perform such an evaluation in purely formal terms because these choices involve practical payoffs. Expertise and the exploitation of information in one area come at the cost of maximal exploitation and expertise elsewhere. Reliance on malleable priors inevitably gives rise to a dilemma between specialization within one particular context and flexibility across a range of different environments. What settles the point at which the visual

[14] We are, in effect, in need of a set of norms of inquiry. While it might seem natural to treat norms of inquiry as a subset of epistemic norms, see Friedman (2020) for an argument that norms of inquiry, what she terms *zetetic norms*, may be systematically in conflict with certain epistemic norms.

system best balances specialization and flexibility depends on the organism's goals in a given context.

The instances of prejudicial bias within the visual system that have been described consist in the preferential selection of information. In the Correll case, the visual system relies on priors, rather than continuing to search for new information in its current environment. In the case of face perception, the development of expertise allows for maximal extraction of information from frequently encountered, own-race faces, to the sacrifice of possible information from less familiar, other-race faces. Formal descriptions of the method by which information, once acquired, is integrated or epistemic norms that exclusively evaluate that process are incapable of identifying or evaluating bias at this prior stage, the stage at which information is acquired.

In fact, this sort of decision cannot be evaluated in purely formal terms that abstract away from the content of the relevant information because the relative value of that content in a given context is the basis on which the decision is made. To evaluate that decision, we need to know how well it serves the various goals of the organism. Nor can we simply plug in a generic epistemic goal such as accuracy or truth. It isn't that one or another decision gets us more or less accuracy so much as different accuracy: Specialization within this face space lets us access accurate information of a certain kind. Specialization within a different face space lets one access information of a different kind. Norms that dictate an attentional strategy relative to a set of facts one seeks to learn do not go far enough. Even what set of facts one should seek to know depends in turn on the goals the inquiry serves.

Successful visual perception does not just involve the acquisition of information per se but the acquisition of the information which best positions us to achieve our goals. Both the goal-directed aspect of visual perception and its capacity to manifest expertise assimilate it to a kind of skill. Thinking of it in these terms offers us an evaluative framework which can accommodate its structural reliance on bias, while picking out instances of the prejudicial kind.

Skilled activity is the capacity to achieve a goal through practice within a given environment. We cannot evaluate a skill except relative to a goal and an environment. There is no absolute standard for assessing tennis skill outside of the parameters provided by conventions that govern the game. Moreover, what constitutes skilled play depends on features of the environment: A competent grass-court player may struggle on a clay surface.

I propose a schema for the evaluation of skilled activity of the following form:

> *Skill*: An individual *S* is skilled at activity Φ relative to a goal *g*, environment *e*, and over timeframe *t* if their practice of Φ in environment *e* within timeframe *t* is likely to position them to achieve goal *g*.
>
> ("is skilled at Φ-ing"$_{<g,t,e>}$ = {x|x is skilled at Φ-ing relative to a goal *g*, in environment *e*, over timeframe *t*.})

This general schema applies to the case of perceptual skill. There is no absolute standard of perceptual skill outside of the epistemic and practical goals we have at any one particular moment. An individual's skill at seeing can only be evaluated relative to a goal, within a given environment and across some set timeframe. The specification of the timeframe is closely tied to the individuation of the environment: How we delineate the timeframe will determine the quantity of change in the environment to which the visual system must adapt. For instance, if the context is construed to include a timeframe during which it grows dark and then light again, a very different perceptual profile will count as skilled than if the timeframe only includes daylight periods.

For my purposes, a crucial upshot of this schema for the evaluation of skill is that it will deliver multiple competing evaluations relative to multiple different goals. What constitutes skilled seeing relative to one goal may not constitute skilled seeing relative to another. In fact, the goals that perception serves can frequently come apart. Even within the subset of an organism's epistemic goals, there will be divergent ends such as accurately representing repeated local stimuli or retaining a sensitivity to unexpected novel items. We can derive competing evaluations of their perceptual skill relative to these different goals. There will be no single verdict on an individual's perceptual skill without an ordering on goals that lets us integrate these competing evaluations. Ultimately, verdicts relative to one or another goal may remain more informative than any such synthesis.

How does this help us identify bias in the cases we have discussed? Just as evaluation of skill has to take place relative to a goal, so the identification of prejudicial bias can only happen relative to a goal. Prejudicial bias limits our capacity to achieve certain goals: More specifically, it is at odds with the goal of a kind of demographic neutrality in our response to others. But a prejudicial bias relative to that goal may facilitate the achievement of other goals.[15]

[15] See Gendler (2011) for a related discussion of the epistemic dilemma that race-based priors give rise to as both encoding them and failing to encode them appear to carry epistemic costs.

We can now define *prejudicial bias*: An instance of formal bias is an instance of prejudicial bias when its effect is to impede a demographically neutral epistemic response to other individuals.

What constitutes such a failure? We do not respond to others in a demographically neutral fashion when their demographic status plays a significant role in determining how we reason or acquire information about them. Responding to others in a demographically neutral epistemic fashion is inconsistent with appeal to base rates that encode social categories. Similarly, it prohibits the adoption of strategies for the extraction of information that vary depending on the social profile of an individual. This leaves much unspecified: What constitutes demographic neutrality will vary depending on context.[16] But the definition at least gives us a handle on the way in which prejudicial bias is identified relative to a goal. We have a multitude of reasons for prizing the suspension of demographic bias in our investigation of and evaluation of others. Some of these are ethical: Doing so promotes the right treatment of others. Some of them are distinctively epistemic: We value neutrality because it preserves our capacity to access information across different contexts, information that is particularly valuable to us when it concerns other people. This criterion for distinguishing prejudicial bias is not formal: To apply it we have to attend to the contents of the relevant states. But prejudicial bias is still an epistemically significant category because the goal in question is an epistemic one: It determines how we gather and use information. We value the capacity to learn about others in ways that disregard their demographic status. It is against this goal that the category of prejudicial bias is defined.

Take the case of face perception. The specialization of one's capacities for facial recognition within a particular area of face space is likely to serve us well relative to the goal of reading and remembering the faces of those in our immediate familial or social circle. But we also value highly the ability to accurately read the faces of other individuals who fall outside of that group. This latter goal falls under the broader aegis of a demographically neutral epistemic response to others. So this specialization, though it constitutes a kind of skilled expertise on one axis of evaluation, is an instance of prejudicial bias.

[16] It might look like this overpredicts instances of prejudicial bias. It identifies prejudicial bias, for instance, when a doctor uses information about demographic categories to predict risk. I accept that these cases are instances of prejudicial bias but instances whose utility relative to other practical and epistemic goals outweighs the disutility involved in the loss of neutrality.

Similarly, a subject's reliance on stereotype-consistent priors may facilitate recognition of a certain range of objects within a stable environment, but it is an impediment to other goals which again fall into the set against which prejudicial bias is defined, namely, the goal of retaining a kind of racial neutrality when perceiving individuals.

Thinking of perception as a skill has other helpful upshots in our endeavor to understand the nature of perceptual bias, both formal and prejudicial. In the first place, doing so moves us toward an understanding of visual perception as an active process that responds dynamically to environmental challenges. In practicing a skill like tennis or cookery we build through repetition a set of mental states capable of appropriately guiding the activity in question.[17] Similarly, by repeatedly seeing a particular environment, we build priors that optimally guide our perception of that environment.

This in turn directs our attention to the role of the context in which the activity is practiced in honing the resultant skill. No matter how gifted an individual practitioner, their skills are inevitably limited by the environment they find themselves in. Practicing running on one surface hones our skills at running on that particular surface. An excessively narrow training environment may limit our capacity to flexibly adapt to a new surface. A bias need not be rooted in an individual performance but in the field the game is played on or even the field the player has consistently practiced on.

Perceptual skill, too, recapitulates the learning environment. One way of avoiding prejudicial visual bias is to ensure that the individual "practices" on an appropriately varied set of samples. But their opportunities to do so may depend significantly on facts about social organization: Practices of segregation, implicit or explicit, limit the learning sample and with it the flexibility of the resulting skill. In doing so, they give rise to manifestations of visual prejudicial bias.[18]

Conclusion

The visual system does not simply respond to a determinate set of information. It selects information on the basis of past exposure and present

[17] This meshes with Stanley and Williamson's (2016) definition of *skill* as "a disposition to form knowledge states appropriate for guiding" the activity in question.

[18] See Munton (2019) for an argument that social structures can cap perceptual skill, in particular via the sensitivity of visual priors to regularities secured by structural injustice.

motivation. That selective process opens the door to varieties of prejudicial bias that we struggle to capture in terms of epistemic analyses designed primarily for evaluating transitions within a fixed body of information. Recognizing the ways in which the visual system is an active, selective process should encourage us to think of it as kind of skill and to adopt an evaluative framework appropriate to its status as such.

Doing so opens the way for us to identify prejudicial bias within a neutrally biased system. That identification must take place relative to a goal. We have a standing goal, of preserving a kind of epistemic neutrality toward others regardless of their demographic properties, which certain cases of bias fall foul of. These are cases of prejudicial bias. Although prejudicial bias is defined in relation to an epistemic goal, the way in which it can arise in the acquisition of information (rather than in aberrant transitions between pieces of information) leaves standard tools of epistemic evaluation poorly placed to identify it within a structurally biased system such as visual perception.

References

Ballantyne, N. (2015). The significance of unpossessed evidence. *The Philosophical Quarterly*, *65*(260), 315–335.

Blum, L. (2004). Stereotypes and stereotyping: A moral analysis. *Philosophical Papers*, *33*(3), 251–289.

BonJour, L. (1985). *The structure of empirical knowledge*. Harvard University Press.

Brownstein, M., & Saul, J. (2016). *Implicit bias and philosophy* (Vols. 1 and 2). Oxford University Press.

Chabris, C. F., & Simons, D. J. (2010). *The invisible gorilla: How our intuitions deceive us*. Random House.

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., & Schwartz, O. (2007). Visual adaptation: Neural, psychological and computational aspects. *Vision Research*, *47*(25), 3125–3131.

Conee, E., & Feldman, R. (1998). The generality problem for reliabilism. *Philosophical Studies*, *89*(1), 1–29.

Conee, E., & Feldman, R. (2004). *Evidentialism: Essays in epistemology*. Oxford University Press.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2007). The influence of stereotypes on decision to shoot. *European Journal of Social Psychology*, *37*(6), 1102–1117.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., and Tracie Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*(6), 1006–1023.

Correll, J., Crawford, M., Wittenbrink, B., & Sadler, M. (2015). Stereotypic vision: How stereotypes disambiguate visual stimuli. *Journal of Personality and Social Psychology*, *108*(2), 219–233.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564.

Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing Black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, *87*(6), 876–893. https://doi.org/10.1037/0022-3514.87.6.876

Elfenbeim, H. A., & Ambady, N. (2002). Is there an in-group advantage in emotion recognition? *Psychological Bulletin*, *128*(2), 243–249.

Fantl, J., & McGrath, M. (2002). Evidence, pragmatics, and justification. *Philosophical Review*, *111*(1), 67–94.

Feldman, J. (2014). Bayesian models of perceptual organization. In J. Wagemans (Ed.), *Oxford handbook of perceptual organization* (pp. 1008–1026). Oxford University Press.

Friedman, J. (2020). The epistemic and the zetetic. *Philosophical Review*, *129*(4), 501–536.

Gauthier, I., Tarr, M., & Bub, D. (2009). *Perceptual expertise: Bridging brain and behavior*. Oxford University Press.

Gendler, T. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, *156*(1), 33–63.

Goldman, A. (1979). What is justified belief? In G. Pappas (Ed.), *Justification and knowledge* (pp. 1–25). D. Reidel.

Harel, A., Kravitz, D., & Baker, C. I. (2013). Beyond perceptual expertise: Revisiting the neural substrates of expert object recognition. *Frontiers in Human Neuroscience*, *7*, Article 885. https://doi.org/10.3389/fnhum.2013.00885

Hawthorne, J. (2003). *Knowledge and lotteries*. Oxford University Press.

Jaquet, E., & Rhodes, G. (2008). Face aftereffects indicate dissociable, but not distinct, coding of male and female faces. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(1), 101–112. https://doi.org/10.1037/0096-1523.34.1.101

Jaquet, E., Rhodes, G., & Hayward, W. G. (2007). Opposite aftereffects for Chinese and Caucasian faces are selective for social category information and not just physical face differences. *The Quarterly Journal of Experimental Psychology*, *60*(11), 1457–1467. https://doi.org/10.1080/17470210701467870

Jaquet, E., Rhodes, G., & Hayward, W. G. (2008). Race-contingent aftereffects suggest distinct perceptual norms for different race faces. *Visual Cognition*, *16*(6), 734–753. https://doi.org/10.1080/13506280701350647

Johnson, G. M. (2020). Algorithmic bias: On the implicit biases of social technology. *Synthese*, *198*(10), 9941–9961.

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, *33*(1), 1–27.

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping and discrimination* (pp. 199–227). Taylor & Francis.

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007). The other-race effect develops during infancy: Evidence of perceptual narrowing. *Psychological Science*, *18*(12), 1084–1089.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*(2), 150–158.

Kornblith, H. (2002). *Knowledge and its place in nature*. Oxford University Press.

Lehrer, K. (1990). *Theory of knowledge*. Routledge.

Little, A. C., DeBruine, L. M., & Jones, B. C. (2005). Sex-contingent face after-effects suggest distinct neural populations code male and female faces. *Proceedings of the*

Royal Society B: Biological Sciences, *272*(1578), 2283–2287. https://doi.org/10.1098/rspb.2005.3220

Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A., & Palumbo, P. (1998). The accuracy and power of sex, social class and ethnic stereotypes: Naturalistic studies in person perception. *Personality and Social Psychology Bulletin*, *24*(12), 1304–1318.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy and Law*, *7*(1), 3–35.

Munton, J. (2019). Perceptual skill and social structure. *Philosophy and Phenomenological Research*, *99*(1), 131–161. https://doi.org/10.1111/phpr.12478

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory and Cognition*, *3*(5), 519–526.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, *81*(2), 181–192.

Payne, B. K., Shimizu, Y., & Jacoby, L. L. (2005). Mental control and visual illusions: Toward explaining race-biased weapon misidentifications. *Journal of Experimental Social Psychology*, *41*(1), 36–47. https://doi.org/10.1016/j.jesp.2004.05.001

Plant, E. A., & Peruche, M. B. (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science*, *16*(3), 180–183.

Rescorla, M. (2015). Bayesian perceptual psychology. In M. Matthen (Ed.), *The Oxford handbook of the philosophy of perception* (pp. 694–716). Oxford University Press.

Scholl, B. J. (2005). Innateness and (Bayesian) visual perception: Reconciling nativism and development. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The structure of the innate mind* (pp. 34–52). Cambridge University Press.

Siegel, S. (2013a). Can selection effects on experience influence its rational role? In T. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 4, pp. 240–270). Oxford University Press.

Siegel, S. (2013b). The epistemic impact of the etiology of experience. *Philosophical Studies*, *162*(3), 697–722.

Siegel, S. (2016). *The rationality of perception*. Oxford University Press.

Stanley, J. (2005). *Knowledge and practical interests*. Oxford University Press.

Stanley, J., & Williamson, T. (2016). Skill. *Noûs*, *51*(4), 713–726.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Science*, *13*(9), 403–409.

Turski, W. G. (1994). *Toward a rationality of emotions: An essay in the philosophy of mind*. Ohio University Press.

Ward, E. J., & Scholl, B. J. (2015). Inattentional blindness reflects limitations on perception, not memory: Evidence from repeated failures of awareness. *Psychonomic Bulletin & Review*, *22*(3), 797–802.

Wells, A., & Matthews, G. (2014). *Attention and emotion: A clinical perspective* (classic ed.). Psychology Press.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*, Article 0058. https://doi.org/10.1038/s41562-017-0058

9

The Trouble of Not Knowing What You Do Not Know

Psychological, Philosophical, and Societal Implications

David Dunning

> The Master said, "Yu, shall I teach you what knowledge is? When you know a thing, to hold that you know it; and when you do not know a thing, to allow that you do not know it;-this is knowledge."
>
> —Confucius (551–479 BCE)

In this quotation, Confucius (2015) makes a sage observation. Any assessment of one's condition, any decision of what course of action to take, requires not one but two judgments. The first is the direct judgment itself, but the second might be more important.

That second judgment is assessing whether one's first judgments should be considered valid or tentative, ones held with certainty or accompanied by doubt. A military general, for example, may design a battle plan that looks like it has the best chance to defeat the enemy, but whether the general goes into battle may depend not so much on the details of the plan but rather whether the general is confident it will succeed. Does the general feel as though every contingency has been accounted for? Is the general sure that the plan is simple enough or instead too complex to execute? Does the plan contain the necessary flexibility in order to weather any number of surprises that might be discovered only after the engagement starts?

In psychology, coming up with a battle plan in the first place can be called the *cognitive task* facing the general. The second task, one of assessing the battle plan's soundness, is the *meta-cognitive task* the general must execute (Dunlosky & Metcalfe, 2009; Metcalfe & Shimamura, 1994). This second

David Dunning, *The Trouble of Not Knowing What You Do Not Know* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press.
DOI: 10.1093/oso/9780197636916.003.0010

task contains many components, but one central component is evaluating the worth of one's reasoning. Is the reasoning accurate, or does it contain errors? Does it neglect important aspects of the task? Does the reasoning flow from complete and accurate information or instead from imperfect data and assumptions? In the example, did the general have all the information needed? Was intelligence about the enemy sound? Are there any unknowns that must be addressed? Does the general have sufficient expertise to draw up a plan? Is their reasoning coherent? Ultimately, should the general be confident or wait instead for another day?

Or, as Confucius put it, is the general wise enough to lead the troops to battle not only having a good plan but knowing it is good?

Recognizing One's Own Ignorance

Here, however, is where Confucius (2015) enters the picture again. Wise knowledge is not only knowing what we know—it also and importantly requires having a reasonable grasp of what we are ignorant of. This chapter is a discussion of how well people have that grasp. Its central contention is that knowing what one doesn't know, having a reasonable understanding of the shape and scope of one's ignorance, makes knowing a hard task. Herein lies the problem not only for the general but for the rest of us as well.

Psychological research over the past few decades has repeatedly demonstrated that people have deep and persistent problems when it comes to recognizing the extent of their ignorance. They often fail to see the line where the solid ground of their knowledge ends and the shifting and dangerous soil of unknowing begins. Real knowledge, the way that Confucius defined it, is something that people have much less than they realize (Dunning, 2005; Dunning et al., 2004).

In this chapter, I discuss the intrinsic difficulty people have in grasping the shape and scope of their ignorance. I discuss three phenomena from empirical psychology—overclaiming knowledge, the illusion of understanding, and the Dunning-Kruger effect—that show that people often claim expertise they in reality do not have, thus revealing their inability to separate their knowledge from their ignorance. I then discuss the reasons why people have such a hard time identifying the boundary between knowledge and ignorance. I describe just how invisible people's ignorance is of them, argue further that people tend to actively neglect what they do not know, and often

suffer from false knowledge that has the look and feel of true knowledge. I end the chapter by discussing both the personal and societal implications of this ignorance of ignorance. At the personal level, people often fail to ask for advice when they need it and fail to recognize true expertise in others. At the societal level, ignorance of ignorance may cause people to dismiss experts on social issues and may leave them vulnerable to the arguments of false prophets bearing enticing but misleading information.

Three Demonstrations Illustrating Ignorance of Ignorance

Many studies in psychology provide ample evidence that people cannot tell where their expertise ends and their ignorance begins. In three different paradigms, researchers easily catch people crossing the boundary between knowledge and ignorance without their ever noticing it.

Overclaiming Knowledge

In our own lab, we have demonstrated this unknowing step into ignorance by giving people simple surveys. We might, for example, give then a list of common financial terms, such as *stock options*, *revolving credit*, and *whole life insurance*, and ask them how familiar they are with the terms (Atir et al., 2015). People show a good deal of familiarity with these terms, which is good.

However, what is interesting is the familiarity they report with a special class of terms, such as *fixed-rate deductions* or *pre-rated stocks*. In one survey, over 90% of respondents reported familiarity with at least one of these items. However, this presents a problem: These items are ones we have simply invented among ourselves in our office. These concepts do not exist. We have made sure they appear nowhere, for example, on the internet. Thus, there is no possible way for respondents to have any familiarity with them.

Yet, respondents report knowledge of and familiarity with items that are impossible to know because these items that do not exist. This phenomenon is known as *overclaiming*, and its existence has been known for many decades in behavioral science (Paulhus et al., 2003). In our studies, participants overclaim knowledge for "concepts" falling into philosophy, biology, and social science (Atir et al., 2015). They report familiarity with fictitious cities such as Cashmere, Oregon, and nonexistent politicians such as Michael Merrington.

Consumers report some acquaintance with food (e.g., Barjolet cheese) and industrial products (e.g., Thompson drill bits) that do not exist (Graeff, 2003).

In a similar vein, survey respondents express opinions about recent governmental actions, such as the International Monetary Act, that have never been proposed (Schuman & Presser, 1980). In 2015, 30% of Republicans versus 19% of Democrats favored bombing the Middle Eastern kingdom of Agrabah—the only problem being that the kingdom existed only in the child's tale of *Aladdin* (Jensen, 2015).

The calendar year 2017 was a banner year for overclaiming in the political realm. In February, 51% of people who had voted for Donald Trump for president the preceding year cited the "Bowling Green Massacre," a completely fictitious event, as a rationale for banning Muslims from entering the United States (England, 2017). In October, two South Carolina legislators proposed a statue to honor Confederate African American soldiers, although history provides no record of any, for obvious reasons (Criss, 2018). Later in December, two Russian radio comedians prompted United Nations Ambassador Nikki Haley to condemn Russian interference in elections taking place in the island nation of Binomo, which existed only in the context of that radio conversation (Mortimer, 2017).

Illusions of Understanding

Overclaiming is not the only way that people demonstrate failures to locate the line between their knowledge and ignorance. Ask people to explain how everyday items work, such as a ballpoint pen or a coat zipper, and they profess confidence in their ability to do so. However, if then asked to go ahead and explain how such items function, people retreat from their confidence to a more humble position. This phenomenon of intellectual confidence that evaporates to humility is known as the *illusion of explanatory depth* (Fisher & Keil, 2016; Rozenblit & Keil, 2002).

The illusion even extends to the political and social arena (Fernbach et al., 2013). Political partisans often proclaim to have a clear idea about how certain social policies work, such as national healthcare or sanctions against a foreign country. However, once asked to explain in detail how these policies operate, partisans realize they do not know as much as they thought. They retreat to a position of more intellectual modesty—and, perhaps more important, moderate their partisanship.

The Dunning-Kruger Effect

The most extensive demonstration, however, that people know not where the geography of their ignorance begins comes from studies on the phenomenon that has come to be known as the *Dunning-Kruger effect* (Dunning, 2011; Kruger & Dunning, 1999). The effect asserts that people of poor expertise fail to recognize just how poor their expertise is. Actually, the theoretical claim is stronger. The claim is that people with poor expertise lack what they need to be able to recognize their shortcomings. It is not that they fail to recognize their deficits; instead, they are simply not in a position to recognize those deficits and should not be expected to do so.

The logic for this claim is rather transparent. No matter whether they are called inexpert, naïve, untrained, or unskilled, people who suffer from the Dunning-Kruger effect simply lack the expertise they need to be able to recognize the expertise they lack. To recognize their deficits requires the exact knowledge they fail to have. As such, they suffer a double curse: Not only are they incompetent but they are also too incompetent to recognize just how deep their incompetence runs. Thus, poor performers often think they are doing just fine when they are doing anything but (Dunning, 2019).

By now, demonstrations of the Dunning-Kruger effect are varied and numerous. People performing badly on tests of logic, grammar, financial literacy, physics, and emotional intelligence all dramatically overestimate how well they think they are doing—often being almost but not quite as positive about their performance as top achievers (Kruger & Dunning, 1999; Sheldon et al., 2014; Williams et al., 2013; for a review, see Dunning, 2011). Students failing or nearly failing an exam in a college course and who in reality are performing at only the 13th percentile relative to their peers think on average that their performance puts them in the 60th to 65th percentile (Dunning et al., 2003; Schlösser et al., 2013). None of this is due to careless judgment on the part of respondents. Offering to pay participants up to $100 for accurate estimates of performance (in this case, on a logical reasoning quiz) does nothing to enhance how accurately respondents report their performances (Ehrlinger et al., 2008).

The Dunning-Kruger effect is also evident outside our laboratory. Gun enthusiasts, participating in a trap-and-skeet shooting competition, who did badly on a quiz about firearm care and safety were just as confident in their skill as were those who did the best on the quiz (Ehrlinger et al., 2008). In medicine, residents completing their rotation in obstetrics/gynecology

at the bottom of their class (and so receive an F on their final exam and a C– for the rotation) think they are achieving a B or B– on the exam and a B+ for the rotation (Edwards et al., 2003). In a similar vein, of 95 first-year medical students learning basic CPR, only three thought they had failed the class when a full 36 did (Vnuk et al., 2006). In more leisurely pursuits, the worst players in debate teams, bridge clubs, chess tournaments, and poker competitions all overestimate how well they are doing (Ehrlinger et al., 2008; Park & Santo-Pinto, 2010; Simons, 2013).

Why Ignorance Is Invisible

Why is human ignorance so invisible to those who possess it? Why do people fail to recognize what they do not know?

Any discussion of this issue must start with a description of the human mind. First, one must concede that what the typical mind knows is truly impressive. By age 60, the typical English speaker knows the equivalent of 48,000 words and their dictionary definitions (Brysbaert et al., 2016). Research subjects can view 2500 photographs, and then distinguish the ones they have seen from photographs taken at a different angle with 87% accuracy—a testament to just how much storage capacity the brain has for detailed memories of visual stimuli (Brady et al., 2008). One cognitive psychologist, looking at the rate at which people add pictures, words, or music to memory, has estimated that the human brain can carry 10^9 bits of learned information over the course of a lifetime, which is more than 50,000 times the text contained in the US Library of Congress (Landauer, 1986).

Moreover, people have a remarkable capacity for taking these bits of information and flexibly applying and combining them to understand the situations they face or to creatively mold new ones (Marsh et al., 2016). Such a capacity to use information to understand situations is essential, in that people never really encounter the exact same situation twice. The cat they see slinking in the bushes is one they have never seen, but they know enough to know it is a cat. Or they need to leave a message outside a friend's house but have no pen or pencil. They know enough to take a piece of charcoal from a nearby grill and write the message on the concrete landing outside their friend's door.

But this capacity for storing, applying, and weaving knowledge—no matter how impressive—pales against the capacity of the world to produce

what people could know. People may know nearly 50,000 words by age 60, but there are over 600,000 word definitions contained in the *Oxford English Dictionary*—with a massive number of other ideas contained in other languages. Just to give one example, there are at least 216 words related to "well-being" in other languages that are untranslatable to English (Lomas, 2016). People may also remember the details of thousands of photographs, but there are an infinite number of images possible on planet Earth. People may know 10^9 bits of information, but that equates to just a few hundred megabytes, well within the capacity of their ordinary desktop computer (Marois & Ivanoff, 2005). It is as Michel de Montaigne (1877) once observed, "there escapes us a hundred times more than comes to our knowledge" and that "if we saw as much of the world as we do not see, we would perceive . . . a perpetual multiplication and vicissitude of forms."

Unknowns Beyond One's Ken

The problem people have is perceiving, or even conceiving of, the multiplicities and vicissitudes that they do not know. In economics, it is customary to place the unknown into three separate categories (Zeckhauser, 2006), two of which are much easier to think about than the third. First, there are things that are unknown, but the probability that they will appear is well defined. As an example, you flip a coin in the air. While in the air, it is unknown whether the coin will land heads or tails, but those two options define a finite and easily decidable set of outcomes (coins rarely, if ever, land and stick on their edge) such that one can assign a 50–50 probability to each outcome. This, according to economists, is a scenario involving risk.

But, second, unknowns can be more unruly. We know the stock market will go up or down over the course of the next year, and this again defines the entire outcome set. However, in this case, we cannot assign exact probabilities to either outcome. This, according to economists, is uncertainty and represents a type of problem that people typically face in the world but which is rarely represented in economic textbooks (Zeckhauser, 2006). People may know which of two teams may win the National Football League's Super Bowl but differ in the specific odds they assign to either team winning.

Unknown Unknowns

Third, unknowns can be even more unruly, in that we fail to recognize all the possible outcomes themselves that may occur. As such, we cannot assign probabilities to them because we simply do not know that they exist. In this, we are in a state of ignorance (Zeckhauser, 2006). We move from the world of unknowns that are known to the murky and complicated world of *unknown unknowns*—risks, events, or even opportunities that are so unknown to us that we do not know that we do not know them.

If known unknowns are questions we do not have the answers to, unknown unknowns are the questions we do not even know we need to ask. Unknown unknowns can importantly shape the fate of human events. Napoleon once set 25,000 soldiers in Haiti to put down a slave rebellion. He was ultimately defeated not by his enemy but by yellow fever, which claimed so many lives that he had to retreat from the island with the 3000 men he had left (Keyes, 2014).

In our work, we have shown that people fail to possess adequate intuitions about what they do not know. For example, consider the word *spontaneity*. How many other English words can you create, from 2 to 11 letters long, from picking and rearranging the letters in this seed word? Some words are easy to see, such as *opt*, *pen*, and *tape*. But how many are there total? It turns out that there are 718 English words that one can generate. However, to the typical person, many of these words remain in the unknown unknown category. These words exist, but a person staring at *spontaneity* has no knowledge of them and thus would never generate them, words such as *pentosan* (a polysaccharide widely distributed in plants), *pontine* (relating to the pons of the brain), or *saponite* (a trioctahedral mineral of the smectite group).

In a word game not unlike this word search, we gave participants puzzles that contained from 100 to 254 solutions. On average, participants thought they had missed 18 possible solutions, but the real figure was 154. In another study, we asked psychology graduate students to go over an experiment and report the methodological flaws it contained. The best students found 69% of the flaws; the worst performers found only 21% of them. . Participants thought their errors of omission mattered—in that once those errors were revealed to them, they lowered the self-ratings of performance. In yet another study involving a word game, they bet less money that they had beaten another student at the game once their errors of omission were pointed out to them (Caputo & Dunning, 2005).

Thus, it appears that people are blind to their errors of omission, presumably because some of those errors lie in the realm of the unknown unknown. In a study of medical malpractice suits, the leading cause could be traced to errors in diagnosis, of which 40% resulted in death and 17% in permanent injury. Of those errors, a majority involved missed diagnoses (54%) rather than delayed (20%) or incorrect (10%) diagnoses (Tehrani et al., 2013). It was errors of omission that proved to be the most common and lethal.

Hypocognition

One important variant of unknown unknowns falls into the category of *hypocognition*. If you do not know what hypocognition is, then you have just experienced it. Hypocognition is lacking a linguistic or cognitive representation for some object, emotion, category, or idea (Wu & Dunning, 2018).

The term was first discussed extensively by anthropologist Robert Levy (1973) in his fieldwork with Tahitians of the Society Islands. During his work, Levy observed that although Tahitians were quite expert in some emotions (in particular, anger), they had no conception of long-term grief. If a person experienced a death in the family, the initial anguish was well understood and public; but any continuing distress seemed unrecognized and unexamined. Islanders would feel grief but not completely understand it, describing themselves instead as feeling "sick" or "strange."

In a similar vein, non-Arabic speakers, when looking at Arabic letters, cluster letters together according to their similarity in a different way than fluent Arabic speakers. Each group experiences the visual nature of the letters differently. Non-speakers just see the physical letter, whereas speakers also internally experience a number of associations, such as sounds and the brushstrokes needed to create those letters. These associations not only change what they experience but make them more accurate and efficient in judging the physical similarity of different letters (Wiley et al., 2016).

Much of human action, thus, might depend not on what people know but rather on what they do not know. Without knowledge of grief, for example, there can be no grief work. In finance, many people remain hypocognitive to the idea of compound interest and so fail to recognize just how much saving money could benefit them or how much debt can penalize them. They simply do not anticipate how much a compounding interest rate can cause any amount of money to grow and grow quickly (Lin et al., 2016). People

ignorant of compound interest also refuse to use financial decision aids, although their use would obviously benefit them (Levy & Tasoff, 2017). In a similar vein, it is estimated that a full third of people suffering from type II diabetes do not know it and do not seek out medical help because they do not recognize that seemingly isolated symptoms (chronic fatigue, numbness, blurry vision, frequent urination) actually indicate a single serious underlying medical condition (Cowie et al., 2006).

Corralling the Unknown

Although so much information lies beyond the ken of the individual human, people could do a lot better dealing with unknowns. The real problem is that not all of the unknown lies in the realm of the unknown unknown but can be brought into the realm of the known unknown or even the known—if people would simply pay more attention to what they might be ignorant of.

That is, people make decisions based on what they know and set aside or do not consider what they do not know, even when knowledge of it is within their grasp. They act as though they have complete information even when information outside their knowledge could change their decision. For example, college students can often remain undecided between a $700 stereo and a $1000 one of somewhat better quality. However, if they are simply reminded that they could also buy $300 worth of music if they bought the cheaper stereo, students quickly break for that option (Frederick et al., 2009).

In a similar vein, people tend to ignore information that is missing, even though it should be relevant to the decision at hand. For example, when buying a car, people fail to ask for the safety record of the car's brand if that record is never mentioned, even though they think it is relevant if mentioned. In addition, when judging the quality of a camera, people make just as confident assessments of the camera's quality if shown only four attributes of the camera as when shown eight (Sanbonmatsu et al., 1992, 2003). People also make more extreme judgments about an object (e.g., a bicycle) when thinking about it in memory, after much information has been forgotten, than when the information is fresh in their mind, again neglecting what they have failed to retain in memory (Sanbonmatsu et al., 1991).

To be sure, people who are expert are better at attending to information that is missing (Sanbonmatsu et al., 1992). Apparently, ridding one's self of hypocognition is useful in aiding people to recognize and weight omissions

in information, thus aiding their choices. Blatantly pointing out to people that there is information they miss and asking them to list what that information is also prompts them to be less overconfident in their decisions (Feduzi & Runde, 2014; Walters et al., 2017).

The Veil of False Belief

The second reason why people fail to recognize the scope of their ignorance is that it is often veiled by beliefs and opinions that are false but which have the look and feel of the truth. Recall that a person learns up to 10^9 bits of information as they navigate life. It would be a surprise, and quite arrogant, to assume that all of those bits of information are accurate. Some of them must be wrong or misleading. This is the reason why, for example, so many people believe Toronto is the capital of Canada (sorry, Ottawa), that the Sahara is the largest desert in the world (actually, the driest spot on the globe is Antarctica), and that Pluto is the farthest planet from the sun (sadly, it has been demoted to a less-than-planetary status). People know enough to come up with an answer to any question posed; those answers do not necessarily have to be true (Tauber et al., 2013).

Consider a survey conducted in my lab the day after the US congressional midterm elections took place in November 2014 (Dunning & Roh, 2018). Participants were asked about their politics and then asked what they believed to be true about social, economic, and political conditions in the country. They were asked, for example, whether the poverty rate had gone down under the administration of President Obama, whether teenage pregnancies were at an all-time high, and whether the budget deficit was shrinking in line with Obama's promises. Half the statements were conservative-friendly; half were more congenial to liberals.

Not surprisingly, conservatives and liberals differed in which facts they claimed were true of the country. Conservatives were more likely to think teenage pregnancies were an epidemic; liberals were more likely to believe that the Obama presidency had lowered the poverty rate. Of key interest, however, 35% to 40% of what each side endorsed as true was actually false (or what they believed to be false was actually true). Teenage pregnancies in 2014 were actually at a several-decade low; the poverty rate had risen in the early days of Obama. In short, conservatives and liberals lived in different factual worlds—and this divergence occurred even though every respondent

on every question had the option of saying "I don't know"; they did not have to guess.

Everyday Paralogia

We have termed this phenomenon of generating false answers *everyday paralogia*, borrowing the terminology from clinical psychology for beliefs that are unsound, illogical, or delusional, and have traced their implications (Dunning & Roh, 2018).

These paralogic beliefs, for example, matter for self-perception. We asked respondents how "well-informed" they considered themselves about civic and national affairs. Not surprisingly, respondents who answered our quiz more correctly also rated themselves as more informed, as they should. Those who opted to say "I don't know" more often rated themselves as less informed—again, appropriate. What was important, however, was the relationship between giving false answers and self-perception. The more respondents gave false answers to our questions, the more they rated themselves as well informed. In short, when it came to a positive self-view, people gave themselves credit for both their right and wrong answers. Clearly, the ability to generate those wrong answers—presumably fed by up to 10^9 bits of material in their brain—hid from people how little they actually knew (Dunning & Roh, 2018; see Dunlosky & Lipko, 2007, for similar effects in self-ratings of text comprehension).

Such divergences in belief mattered for behavior as well. In a follow-up analysis, we examined how sensitive respondents' true–false responses were (a) to the truth or (b) to their partisan leanings, in that they tended to endorse statements friendly to their politics as true. Then we looked at who reported they had voted the day before. Voters, relative to non-voters, were not clearly more sensitive to the truth in their responses. However, respondents whose responses were sensitive to their partisan leanings were much more likely to vote than those whose answers were more even-handed (Dunning & Roh, 2018).

The pattern of people giving themselves positive credit for paralogic answers is one we have seen in a number of areas—from civics to beliefs about financial literacy. Those who think they know something that is wrong are as confident in their ability as those giving right answers. Consider the following financial question: You invest $100 in a bond that pays 6% per year,

compounded annually. If you leave it untouched, how long will it take for the value of the bond to double?

If you said 12 years, you have a good handle on compound interest and exponential growth. That is the correct answer. Many people, however, fail to understand how compound interest works and believe, instead, that investments grow in a more incremental and linear manner. They are likely to say it will take at least 16 years for your investment to double. In that, they are wrong, but people who consistently follow a more "linear" model of investment are likely to be as certain in their answers as are their peers who are consistently right (Williams et al., 2013).

Application to Beginners

This ability to form paralogical beliefs also explains how the Dunning-Kruger effect tends to be so widespread among beginners at a task. People do not necessarily approach new tasks with overconfidence. However, give them just a modicum of experience and feedback, and they display a burst of confidence that far outstrips any level of performance they achieve at the task. Put differently, a little learning might not necessarily be a dangerous thing, but it is a thing that feeds undue confidence in one's ability.

In our lab, we have demonstrated this beginner's bubble of overconfidence by asking people to tackle a computer game in which they must diagnose who is infected with a zombie disease in a post-apocalyptic world (Sanchez & Dunning, 2018). At the very beginning of the task, participants are quite cautious in the evaluations they make of their diagnoses, but it only takes a few diagnoses before their confidence explodes into something running far ahead of their accuracy rate. After a while, this inflation of confidence cools down considerably, but accuracy never catches up to it.

Further, we have found that the burst occurs because people take small scraps of experience and feedback as they begin the task and spin elaborate and self-assured theories about how to diagnose zombie diseases. In effect, they give their early experience too much credit, in that any small sample of experience is likely to contain a lot of noise and misleading signals (Sanchez & Dunning, 2018). It is only after extensive experience that one can begin to separate true patterns from chaotic noise. People learn this fact only after a while, while the degree of their overconfidence cools off; but they never completely learn the lesson that the information they have in their head, and their

creativity at weaving those pieces together into plausible theories, makes it much too easy to reach an answer to any question, just not necessarily the right answer.

Implications for Self and Society

Naturally, all these difficulties in identifying the shape and scope of one's ignorance carry many implications for human life. Some of them are personal and affect the individual. Some of them ultimately affect society at large. Many of them, interestingly, intersect with philosophical topics and issues.

Self

Take the issue, for example, of rational ignorance (Downs, 1957). It is impossible for a person to know everything there is to know about the world. Much too much information now exists to allow a person to become a contemporary renaissance person. Further, it is imprudent to try to be that renaissance person, in that it is just not worth the effort to become expert in some topics. My life is not diminished, for example, by not knowing the history of the Baltic States or basic ballistics or the poems of Middle English. As such, one can make the argument that there are clear circumstances where it is rational to be ignorant of a topic, times when the person is better off being uninformed than expending the effort to become well informed.

No doubt, there is some truth to this argument, but the invisibility of personal ignorance makes it intrinsically difficult to know whether one is achieving rational ignorance or not. Essentially, blind to an area, we do not know enough to know if we are better off being blind. To make such a proper call about whether we can remain uneducated about an area, we would have to put in the effort of gaining an education in that area. As such, we are not in a position to make a call about whether our ignorance is adaptive or maladaptive.

To be sure, in some areas, one could imagine the call would be easy to make. My work would likely not improve if I spent a few years studying the science and aesthetics of different typefaces. But how does one really know? For example, in my line of psychology, it would seem clear that I would never have to learn Bose-Einstein statistics, an esoteric strand of analysis developed

in physics to account for the quantum behavior of small particles in various energy states. However, such a call might be preliminary and wrong—in that other psychological researchers have begun successfully applying such statistics to the errors people make in everyday decisions (Busemeyer et al., 2011).

Asking for Advice

Beyond deciding when it might be appropriate to be ignorant, people often face a much simpler question: Are they ignorant, and should they turn to the expertise of others. It is inefficient for each person to know or try to learn everything. A society benefits if people trade off which areas they will be expert in so that everyone can gain access to necessary expertise when they need it.

The problem is that the invisibility of ignorance suggests that people have substantial trouble knowing when they need to rely on someone else's expertise rather than their own. Any academic instructor who has had student advisees has seen the syndrome: Students with the most need of advice never come into the office to get it until it is too late. Work on in our lab demonstrates this *advisor's paradox*, in that people with the most need for advice are no more likely to seek it out than those who do not need it.

In these studies, we give participants a quiz that should be of some interest and familiarity to them. For example, it may be a survey about household hazards given to parents with at least one child under the age of 6 at home. They are given small monetary bonuses for each correct answer, but they can also ask for advice—to see how another person has responded to the same question (with no guarantee that this other person got the question right)—for a reduced bonus. We find that respondents rarely ask for advice (13% of the time for the household safety questionnaire) and that poor performers ask for no more advice than do top performers (Yan & Dunning, 2018).

Judging Others

There is a further problem, however, that follows from the Dunning-Kruger effect: Flawed expertise not only prevents people from seeing their own ignorance; it also prevents them from recognizing superior competence in others. We have termed this problem the *Cassandra quandary*, after the Greek myth of the princess who was given the gift of true prophecy but also cursed by

the gods to be disbelieved by all other humans. This inability to evaluate the expertise of others is more severe among poor performers, who cannot accurately identify which individuals are best to approach for advice. In one demonstration of this, we gave participants the answers that other people had given to a quiz on financial literacy. We asked who they would want to approach for financial advice. Of those with perfect scores on the quiz themselves, 91% chose the other person with a perfect score. Only 9% of those posting the worst score on the quiz chose similarly (Dunning & Cone, 2018).

Society

This theme of expertise carries over when thinking about societal implications. If people cannot identify either the expert or the worth of their expertise, society will be worse off in the aggregate.

Dismissing Expertise

The invisibility of ignorance promotes problems in dealing with expertise in two ways, which on the surface may seem contradictory. On the one hand, not being able to identify expertise may cause the less able to dismiss or discount it. At an extreme, nonexperts may mistakenly think their knowledge rivals that of experts. One sees signs of this in disparate areas of human life. A recent Pew survey showed that 83% of Americans felt they "understood the challenges police face on the job," but only 13% of police officers agreed that citizens had sufficient understanding. It is likely that the public does not know how often the police have to deal with verbal abuse (two-thirds of officers reported experiencing such abuse within the last month) or how often they enter situations of uncertain safety (Morin et al., 2017).

Or, in the area of medicine, people may substitute their judgment for that of the doctor, an action called *epistemic trespassing* (Ballantyne, 2019). A study that examined 50 years of data revealed that patients failed to substantively adhere to their doctor's instructions in roughly 45% of cases, depending mostly on the specific malady in question (DiMatteo, 2004). In the case of high chronic blood pressure, misbelief and ignorance are the

main reasons people fail to adhere to the regimen their physician gives them. Patients may misunderstand the rationale for the treatment, misunderstand their condition, or overestimate their skill at controlling their blood pressure without medication (Kirscht & Rosenstock, 1977; Patel & Taylor, 2002; Svensson et al., 2000).

Ignorance and misunderstanding may also underlie, at least in part, distrust of scientific expertise. A significant proportion of the American public distrusts scientific conclusions, for example, about climate change and the safety of vaccines. Underlying this distrust is a belief that scientists are free to conclude and say whatever they wish to, but every scientist knows that is not the case. To make a statement to scientific standards, one must follow the strict and constraining rules of the scientific method and generate data that support the scientist's conclusion. If the data fail to cooperate, the scientist cannot make the claim.

In our work, we have shown that people who believe scientists can say whatever they please fail to show adequate knowledge of the scientific method, its constraining rules, and how those rules limit scientific claims. In short, hypocognitive of the scientific method, people misunderstand the basis for scientific claims. Those with fuller knowledge of scientific rules and their constraints turn out to trust scientists more (De Oliveira Chen & Dunning, 2018).

Gullibility to False Information

If the invisibility of ignorance leads to improper dismissal of true and valuable expertise, it also creates a cross-cutting problem—a gullibility to false information that purports to be fact or expertise. Upon its release, a deeply flawed 2014 study linking water fluoridation to lower IQs in children generated Twitter views and shares numbering in the tens of thousands (Vogel, 2017). The National Science Foundation has recently had to publicly deny a report that it was running a child prostitution ring on Mars (Holley, 2017).

The human inability to tell true from false but plausible information may lead people to be too credulous to such fake news and false information. Actually, the psychological situation may be worse than that. People show a *truth bias*, assuming that any new information they encounter is more likely to be true than false (DePaulo et al., 1997; Vrij, 2000; Zuckerman et al., 1981).

Such a bias is reasonable for human activities like conversation—imagine a world in which people routinely disbelieved every single thing other people told them while chatting—but a truth bias may lead to costly gullibility in a digital world where the other person (if it is another person) shows no concern about maintaining honesty.

Concluding Remarks

> Better light a candle than curse the darkness.
>
> —William L. Watkinson (2009)

I began this chapter with Confucius. Let me end with the sermons of William L. Watkinson. As Confucius initially pointed out, real knowledge involves not only awareness of a fact but also the confidence to act on that fact. Much of this confidence depends on how people evaluate their expertise, but this chapter has pointed out all the ways in which that evaluation can go awry. Knowing what you know intrinsically requires also knowing what you do not know, but the problem here is the inherent invisibility of ignorance. People often fail to have a correct and authentic understanding of the shape and scope of what they do not know.

I have taken as my mission in this chapter to light a candle and expose just how difficult it is to see one's own ignorance. The candle reveals just how much darkness there is to curse. A reader might think exposing all this darkness may not be worth it, but I believe the both the old master and Pastor Watkinson would disagree. As Confucius himself further admonished, when you have faults, do not fear to abandon them. Learning how difficult it is to recognize one's ignorance may be an important initial step toward achieving knowledge as Confucius defined it. Lighting that candle, thus, may be an essential first step for finding the way.

Acknowledgment

The writing of this chapter was supported by a grant from the Intellectual Humility in Public Discourse Project at the University of Connecticut, underwritten by the Templeton Foundation.

References

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, *26*(8), 1295–1303.

Ballantyne, N. (2019). Epistemic trespassing. *Mind*, *128*(510), 367–395. https://doi.org/10.1093/mind/fzx042

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America 105*(38), 14325–14329.

Brysbaert, M., Stevens, S., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, Article 1116.

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*(2), 193–218.

Caputo, D., & Dunning, D. (2005). What you don't know: The role played by errors of omission in imperfect self-assessments. *Journal of Experimental Social Psychology*, *41*(5), 488–505.

Confucius. (2015). *The analects of Confucius* (R. Eno, trans.). Bloomington, IN: IUScholar Works.

Cowie, C. C., Rust, K. F., Byrd-Holt, D. D., Eberhardt, M. S., Flegal, K. M., Engelgau, M. M., Saydah, S. H., Williams, D. E., Geiss, L. S., & Gregg, E. W. (2006). Prevalence of diabetes and impaired fasting glucose in adults in the U.S. population. *Diabetes Care*, *29*(6), 1263–1268.

Criss, D. (2018, January 4). *Republicans in South Carolina want to honor black Confederate soldiers. There's just one problem . . .* CNN. https://www.cnn.com/2018/01/03/us/black-confederate-monument-trnd/index.html

de Montaigne, M. (1877). *Essays* (Vol. 16, C. Cotton, Trans.). Project Gutenberg.

De Oliveira Chen, S., & Dunning, D. (2018). *Knowledge of the scientific method and trust in scientists* [Unpublished manuscript]. University of Michigan.

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy–confidence correlation in the detection of deception. *Personality and Social Psychology Review*, *1*(4), 346–357.

DiMatteo, M. R. (2004). Variations in patient's adherence to medical recommendations: A quantitative review of 50 years of research. *Medical Care*, *42*(3), 200–209.

Downs, A. (1957). An economic theory of political action in a democracy. *Journal of Political Economy*, *65*(2), 135–150.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*(4), 228–232.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. SAGE Publications.

Dunning, D. (2005). *Self-insight: Roadblocks and detours on the path to knowing thyself*. Psychology Press.

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In J. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 247–296). Elsevier.

Dunning, D. (2019). Self and the best option illusion. *Self and Identity*, *18*(4), 349–362.

Dunning, D., & Cone, J. (2018). *The Cassandra quandary: How imperfect expertise prevents people from recognizing superior performance in others* [Unpublished manuscript]. University of Michigan.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*(3), 71–106.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83–86.

Dunning, D., & Roh, S. (2018). *Everyday paralogia: Mistaken beliefs bolster a false sense of expertise* [Unpublished manuscript]. University of Michigan.

Edwards, R. K., Kellner, K. R., Sistrom, C. L., & Magyari, E. J. (2003). Medical student self-assessment of performance on an obstetrics and gynecology clerkship. *American Journal of Obstetrics and Gynecology*, *188*(4), 1078–1082.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware? Further explorations of (lack of) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121.

England, C. (2017, February 10). Most Donald Trump supporters still think the "Bowling Green massacre" is real. *The Independent*. https://www.independent.co.uk/news/world/americas/donald-trump-supporters-bowling-green-massacre-real-kellyanne-conway-misspoke-masterminds-white-a7573701.html

Feduzi, A., & Runde, J. (2014). Uncovering unknown unknowns: Towards a Baconian approach to management decision-making. *Organizational Behavior and Human Decision Processes*, *124*(2), 268–283.

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*(6), 939–946.

Fisher, M., & Keil, F. C. (2016). The curse of expertise: When more knowledge leads to miscalibrated explanatory insight. *Cognitive Science*, *40*(5), 1251–1269.

Frederick, S., Novemsky, N., Wang, J., Dhar, R., & Nowlis, S. (2009). Opportunity cost neglect. *Journal of Consumer Research*, *36*(4), 553–561.

Graeff, T. R. (2003). Exploring consumers' answers to survey questions: Are uninformed responses truly uninformed? *Psychology and Marketing*, *20*(7), 643–667.

Holley, P. (2017, July 1). No, NASA is not hiding kidnapped children on Mars. *Washington Post*. https://www.washingtonpost.com/news/speaking-of-science/wp/2017/07/01/no-alex-jones-nasa-is-not-hiding-kidnapped-children-on-mars-nasa-says/?utm_term=.b8f150d0ec1c

Jensen, T. (2015, December 18). *Trump lead grows nationally; 41% of his voters want to bomb country from Aladdin; Clinton maintains big lead*. Public Policy Polling. https://www.publicpolicypolling.com/polls/trump-lead-grows-nationally-41-of-his-voters-want-to-bomb-country-from-aladdin-clinton-maintains-big-lead/

Keyes, P. (2014, July 8). *Yellow fever: Napoleon's most formidable opponent*. Historia Obscura. http://www.historiaobscura.com/yellow-fever-napoleons-most-formidable-opponent/

Kirscht, J. P., & Rosenstock, I. M. (1977). Patient adherence to antihypertension medical regimens. *Journal of Community Health*, *3*(2), 115–124.

Kruger, J. M., & 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.

Landauer, T. K. (1986). How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, *10*(4), 477–493.

Levy, M. R., & Tasoff, J. (2017). Exponential-growth bias and overconfidence. *Journal of Economic Psychology*, *58*, 1–14.

Levy, R. I. (1973). *Tahitians: Mind and experience in the Society Islands*. University of Chicago Press.

Lin, J. T., Bumcrot, C., Ulicny, T., Lusardi, A., Mottola, G., Kieffer, C., & Walsh, G. (2016). *Financial capability in the United States 2016*. FINRA Investor Education Foundation.

Lomas, T. (2016). Towards a positive cross-cultural lexicography: Enriching our emotional landscape through 216 "untranslatable" words pertaining to well-being. *The Journal of Positive Psychology*, *11*(5), 546–558.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, *9*(6), 296–305.

Marsh, E. J., Cantor, A., & Brashier, N. (2016). Believing that humans swallow spiders in their sleep: False beliefs as side effects of the processes that support accurate knowledge. *Psychology of Learning and Motivation*, *64*, 93–132.

Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing*. MIT Press.

Morin, R., Parker, K., Stepler, R., & Mercer, A. (2017). *Behind the badge: Amid protests and calls for reform, how police view their jobs, key issues and recent fatal encounters between blacks and police*. Pew Research Center.

Mortimer, C. (2017, December 29). Nikki Haley seemingly tricked by Russian pranksters into commenting on fictional country "Bimono." *The Independent*. https://www.independent.co.uk/news/world/americas/us-politics/nikki-haley-russia-binomo-prank-us-ambassador-election-hacking-interfere-putin-a8133041.html

Park, Y. J., & Santos-Pinto, L. (2010). Overconfidence in tournaments: Evidence from the field. *Theory and Decision*, *69*(1), 143–166.

Patel, R. P., & Taylor, S. D. (2002). Factors affecting medication adherence in hypertensive patients. *Annals of Pharmacotherapy*, *36*(1), 40–45.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*(4), 890–904.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.

Sanbonmatsu, D. M., Kardes, F. R., & Herr, P. M. (1992). The role of prior knowledge and missing information in multiattribute evaluation. *Organizational Behavior and Human Decision Processes*, *51*(1), 76–91.

Sanbonmatsu, D. M., Kardes, F. R., Houghton, D. C., Ho, E. A., & Posavac, S. S. (2003). Overestimating the importance of the given information in multiattribute consumer judgment. *Journal of Consumer Psychology*, *13*(3), 289–300.

Sanbonmatsu, D. M., Kardes, F. R., & Sansone, C. (1991). Remembering less and inferring more: The effects of the timing of judgment on inferences about unknown attributes. *Journal of Personality and Social Psychology*, *61*(4), 546–554.

Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, *114*(1), 10–28.

Schlösser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How unaware are the unskilled? Empirical tests of the "signal extraction" counterexplanation for the

Dunning-Kruger effect in self-evaluations of performance. *Journal of Economic Psychology*, *39*, 85–100.
Schuman, H., & Presser, S. (1980). Public opinion and public ignorance: The fine line between attitudes and nonattitudes. *American Journal of Sociology*, *85*(5), 1214–1225.
Sheldon, O. J., Dunning, D., & Ames, D. R. (2014). Emotionally unskilled, unaware, and uninterested in learning more: Biased self-assessments of emotional intelligence. *Journal of Applied Psychology*, *99*(1), 125–137.
Simons, D. J. (2013). Unskilled and optimistic: Overconfident predictions despite calibrated knowledge of relative skill. *Psychonomic Bulletin and Review*, *20*(3), 601–607.
Svensson, S., Kjellgren, K. I., Ahlner, J., & Säljö, R. (2000). Reasons for adherence with antihypertensive medication. *International Journal of Cardiology*, *76*(2–3), 157–163.
Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavioral Research*, *45*(4), 1115–1143.
Tehrani, A., Saber, S., Lee, H., Mathews, S. C., Shore, A., Makary, M. A., Pronovost, P. J., & Newman-Toker, D. E. (2013). 25-year summary of US malpractice claims for diagnostic errors 1986–2010: An analysis from the National Practitioner Data Bank. *British Medical Journal Quality and Safety*, *22*(8), 672–680.
Vnuk, A., Owen, H., & Plummer, J. (2006). Assessing proficiency in adult basic life support: Student and expert assessment and the impact of video recording. *Medical Teacher*, *28*(5), 429–434.
Vogel, L. (2017). Viral misinformation threatens public health. *CMAJ 189*(50), Article E1567.
Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and implications for professional practice*. John Wiley & Sons.
Walters, D. J., Fernbach, P. M., Fox, C. R., & Sloman, S. A. (2017). Known unknowns: A critical determinant of confidence and calibration. *Management Science*, *63*(12), 4298–4307.
Watkinson, W. L. (2009). *The supreme conquest, and other sermons preached in America*. New York: Bibliolife.
Wiley, R. W., Wilson, C., & Rapp, B. (2016). The effects of alphabet and expertise on letter perception. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(8), 1186–1203.
Williams, E. F., Dunning, D., & Kruger, J. (2013). The hobgoblin of consistency: Algorithmic judgment strategies underlie inflated self-assessments of performance. *Journal of Personality and Social Psychology*, *104*(6), 976–994.
Wu, K., & Dunning, D. (2018). Hypocognition: Making sense of the world beyond one's conceptual reach. *Review of General Psychology*, *22*(1), 25–35.
Yan, H., & Dunning, D. (2018). *Advice-seeking among poor performers: A behavioral consequence of the Dunning-Kruger effect* [Unpublished manuscript]. University of Michigan.
Zeckhauser, R. J. (2006). Investing in the unknown and unknowable. *Capitalism and Society*, *1*(2), 1–41.
Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 14, pp. 1–57). Academic Press.

10

Novices and Expert Disagreement

Nathan Ballantyne[*]

I always follow orders when they make sense.
—Colonel John Paul Stapp (1910–1999), US Air Force medical doctor, biophysicist, and rocket sled test driver

When a true genius appears in the world you may know him by this sign; that the dunces are all in confederacy against him.
—Jonathan Swift, *Thoughts on Various Subjects, Moral and Diverting* (1706)

We know little about the world without trusting experts. Experts themselves are in a similar predicament because the proliferation and growth of expert knowledge requires experts to trust other experts. But when we defer to experts, we sometimes encounter a serious challenge: expert disagreement. We novices need to determine which experts to trust. That challenge is the focus of this chapter. Here are three examples to focus the discussion.

One recent meta-study of climate science research shows that 97% of climate scientists accept that climate change is caused by human activity. There is a consensus about anthropogenic global warming (Cook et al. 2013, 2016). Although I have watched my fair share of documentaries and have read scores of popular-science articles on climate change, I am mostly ignorant about the empirical facts and theoretical models behind the consensus. I am neither a climate scientist nor a student of climate science. And yet I hold a view about the matter by deferring to the majority of the experts. A bit more

[*] Material in this chapter is drawn from Nathan Ballantyne's *Knowing Our Limits* (Oxford University Press, 2019) and is reproduced with permission of Oxford University Press through PLSclear.

Nathan Ballantyne, *Novices and Expert Disagreement* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press. © Oxford University Press 2022.
DOI: 10.1093/oso/9780197636916.003.0011

specifically, I defer to the majority on the basis of my understanding of how science works, the distribution of expert opinion, and the financial and political pressures that have encouraged some scientists to dissent from the consensus.

As a second example, here's a story about an unnamed friend of mine. After the general election in the United States in November 2016, my friend started reading about foreign policy and international relations. He told me he wanted to understand the potential implications of an American withdrawal from the North Atlantic Treaty Organization and other pertinent questions concerning the threat of nuclear war. Touched a little by obsession, he devoured articles and analyses, finding sharp disagreements among experts. Insofar as it was possible, he checked the credentials and forecasting track records of the rival experts. But he said he couldn't really gauge the credibility of most experts in order to sift out the most trustworthy ones. A few weeks after the presidential inauguration in January 2017, I asked him what he thought now. He said he wasn't sure what to make of the complex issues, but he told me he was ready to quit his foreign policy reading and get on with the task of constructing a fallout shelter in his backyard.

A third example: In the 1970s and 1980s, Linus Pauling, the Nobel Prize–winning chemist, asserted that mega-doses of vitamin C could effectively treat cancer. Pauling was dismissed by the medical community: High-dose vitamin C therapies didn't work. But many cancer patients and their families knew about Pauling's view and urged doctors to prescribe mega-doses of vitamin C. One oncologist remembered how, during Pauling's heyday as supplement guru, families would pressure doctors to prescribe mega-vitamins. "We struggled with that," the oncologist recalled. "They would say, 'Doctor, do you have a Nobel Prize?' "[1] Patients and families knew that doctors and scientists didn't agree, so they deferred to the celebrated Pauling.

Those examples are three among many, but they illustrate our main problem. It is easy to find disagreement among experts on many important issues. All of us are novices about the vast majority of questions, so the problem is ours. How should novices react to finding out that experts disagree? And when is it reasonable for novices to defer to one side? This is what I call the *problem of conflicting expert testimony*. Given that many of our

[1] Offit (2013, p. 55) quotes a pediatric oncologist, Dr. John M. Maris, chief of oncology and director of the Center for Childhood Cancer Research at the Children's Hospital of Philadelphia.

controversial opinions rely in some way on testimony from experts, we can only evaluate our views by grappling with this problem.

A few notes on terminology are in order. I will stipulate that an *epistemic expert* about a question has sufficient evidence and skills needed to answer that question reliably. Although experts have some relatively high degree of epistemic competence, they need not be infallible. Let's say that a *novice* about a question lacks the relevant sufficient evidence and skills to answer the question reliably on their own. Even though novices lack expertise, they can still have reliably formed views, so long as they can defer to experts who have reliably formed views. *Deference* is simply a matter of believing a putative expert's testimony. The idea of putative expertise is important because novices can defer to others who are also novices but present themselves as experts. Deference may be reasonable or unreasonable. If a novice has sufficient reason to believe that a putative expert is a trustworthy source of information on a question and takes the expert at their word, then the novice defers *reasonably*. On the other hand, if a novice lacks reason to trust a putative expert but still believes the expert's testimony, the novice defers *unreasonably*. Unreasonable deference involves an epistemic shortcoming.

The problem of conflicting expert testimony is perennial. Plato discussed it in his dialogues (LaBarge, 1997 and Hardy, 2010). Augustine of Hippo, in his discourse "The Advantage of Believing," touched upon one crucial aspect of the problem when he asked, "[H]ow will we fools be able to find a wise man?" (391–392/1947, p. 429, 13.28). Augustine noted that most people will recognize that fools are better off obeying the precepts of the wise than living according to their own judgments (391–392/1947, p. 428, 12.27). But from the fool's perspective, the right advisors don't leap out. Augustine suggests the fool will be unable to pick out the wise person from among all the fakers and the frauds because the fool doesn't know wisdom in themselves and, thus, can't recognize it in others (391–392/1947, p. 429, 13.28).

When faced with the problem of conflicting expert testimony, what a novice needs to know is this: In situations where I confront expert disagreement, what must I do to respond reasonably, in general, and to defer reasonably, in particular? That issue has only been obliquely addressed in discussions. Philosophers have, in the main, investigated whether it's possible, in principle, for a novice to defer reasonably to one expert. They have considered different kinds of empirical evidence a novice could use to assess putative expertise. But proving in the abstract that reasonable deference is possible is not too helpful when we need to know whether deference is

reasonable in our own case. To shift our perspective on the problem, I propose that we think of it as a matter of cognitive regulation.[2] The novice needs guidance to manage conflicting expert testimony more effectively in order to figure out when and to whom deference is reasonable.

Here I make a start on developing some guidance by addressing a pair of questions. First, what is reasonable deference? Or, in other words, what is required for a novice to defer reasonably to one side in a conflict between experts? Second, is reasonable deference easy or hard? That is, how difficult is it for novices to satisfy the conditions for reasonable deference? As I proceed, I offer an account of reasonable deference that will help novices know how to respond to learning about expert disagreement. Presumably, if novices know the conditions for reasonable deference and know when those conditions are satisfied in ordinary situations, they will be well positioned to regulate their deference. Then I explain why reasonable deference is so difficult by considering the psychology of perceiving expertise as well as the social conditions that produce misinformation about experts. The cognitive mechanisms and social world that influence novices' judgments about experts are highly relevant to the problem of conflicting expert testimony, but this fact has been neglected in ongoing discussions. I conclude with a few observations about the implications of the account for both novices and experts.

Before I start, I want to underline how the problem raises weighty questions about our social and political commitments. In a liberal democracy, for example, novices are expected to be judges of facts and values. Novices vote, serve on juries, and hold political office. They are confronted by expert disagreement. They decide to defer to one side or perhaps go it alone. But when we embrace the ambitions of liberal democracy, we will want to promote the possibility of reasonable deference in important cases. We want novices to use their own autonomous judgments, constrained by some normative conditions, as their basis for deference to experts. We don't want the manipulations of corporations, the media, or authoritarian rule turning voters, juries, and elected officials into marionettes. And many people don't want the experts to decide on behalf of the novices while at the same time hoping novices will actually defer to the right experts. Reasonable deference promises to help novices balance the competing values of autonomy and trust. But threats to reasonable deference are threats to the ideals of liberal democracy.

[2] For more on epistemological theorizing that aims at cognitive regulation, see Ballantyne (2019).

Institutions of higher education in a liberal democracy are sometimes presumed to help novices learn the art of reasonable deference. At Oxford University, just a few years before the First World War, a philosophy professor, John Alexander Smith, made a promise to his students:

> All of you, gentlemen, will have different careers—some of you will be lawyers, some of you will be soldiers, some will be doctors or engineers, some will be government servants, some will be landowners or politicians. Let me tell you at once that nothing I say during these lecturers will be of the slightest use to you in any of the fields in which you will attempt to exercise your skills. But one thing I can promise you: if you continue with this course of lectures to the end, you will always be able to know when men are talking rot.[3]

About 100 years later, Andrew Delbanco, an American studies professor at Columbia University, wrote, "[T]he best chance we have to maintain a functioning democracy is a citizenry that can tell the difference between demagoguery and responsible arguments. . . . [T]he most important thing one can acquire in college is a well-functioning bullshit meter. It's a technology that will never become obsolete" (2012, p. 29). Academics, administrators, and benefactors lap this sort of stuff up. And maybe claims like these are even true. But our highest social and educational ideals may be undermined by epistemological reflection. What if reasonable deference is rarely feasible for the great majority of novices? What if novices' BS meters do not always work as well as we might wish they did? And what if a college or university education—even a really expensive one—can't provide a reliable BS meter?[4]

[3] The source of this quotation is indirect and thus uncertain. Before the First World War, Smith gave a lecture at Oxford, and Harold Macmillan was in the audience. Macmillan (who later served as the British prime minister) reported Smith's words to Isaiah Berlin, who in turn reported them to an interviewer. See Berlin and Jahanbegloo (1992, p. 29).

[4] If a college education fails to equip students with a well-functioning BS meter (see Arum & Roksa, 2011), what other options are there? Students may try the school of hard knocks. The late chef and writer Anthony Bourdain remarked in an interview:

> I was a long-time drug addict, and one of the things drug addiction did, especially when you have to score cocaine or heroin every day on the streets of New York—you learn a lot of skills that are useful when dealing with Hollywood or the business world. In a world full of bullshit, when you need something as badly as drugs, your bullshit detector gets pretty acute. Can I trust this guy with money? Is this guy's package going to be all he says it was? It makes it a lot easier to navigate your way through Hollywood when you find yourself at a table and everybody says, "We're all big fans of your work." . . . You don't fall victim to amateur bullshit when you've put up with professional bullshit (Woods, 2014).

The problem of conflicting expert testimony forces us to consider uncomfortable possibilities. Here are just two. First, if reasonable deference is a vain hope for many novices, their controversial beliefs face a significant threat. Arguably, novices won't be able to regard their beliefs as reasonable unless they can defer reasonably. Second, insofar as reasonable deference is rare, some observers may rethink their social, political, and educational ideals.

Speaking for myself, more widespread reasonable deference is worth aspiring to as a goal in our communities and in our own intellectual lives. Even if it is a lofty goal, we should not give up on it. So let's begin by asking: What is it?

What Is Reasonable Deference?

The idea of reasonable deference is an important, but curiously neglected, element of the problem of conflicting expert testimony. To a first approximation, novices can defer reasonably when they have sufficient reason to believe one expert over another. But that's only a first pass. To get a better grip on the idea, I will explore an example in which critical background details are stipulated, letting me simplify what is often complex about a novice's deference to an expert and then isolate the essential features of reasonable deference.

You are flying a small aircraft through stormy skies at night. You are navigating exclusively by radio. Your destination is an airstrip on a small island, far from the mainland. To ensure you are on the right course, you need some information. Running low on fuel, you know you must correct course soon if the winds have swayed you even one degree from your original flight plan. But tonight your navigational instruments are not working. So you dispatch a radio call to air traffic controllers in the region, requesting further guidance. In response, you receive advice from two air traffic controllers. Their advice conflicts. After you curse under your breath and get over your initial sense of disbelief, you begin to wonder, which expert should you trust, if either one? What explains the conflict here?[5]

[5] Thanks to Noah Hahn for informing me that, for logistical and legal reasons, two air traffic controllers are typically never assigned to guide one aircraft. But on rare occasion pilots may encounter "rogue" radio transmissions, sent by hoaxers with VHF radios dialed to official frequencies. These radio pirates sometimes imitate genuine controllers' messages, endangering pilots and panicking the real controllers. For example, during several weeks in 1993, an unemployed custodian in Roanoke, Virginia, regularly drove around the local airport in his Buick, dispatching misleading messages on his transmitter. He told pilots their runway was closed, that they weren't cleared to land, and that they needed to switch radio frequencies. He sang a line from the horror movie *Child's Play 3*.

Let's assume that you have sufficient reason to eliminate one kind of explanation for the apparent conflict: that the two controllers do not actually disagree. They might only apparently disagree because one uses nautical miles and the other uses land miles or because you misheard the radio communications and mistakenly think the advice conflicts. But you've triple-checked and you have excellent reason to believe the conflict is genuine, in the sense that the experts give incompatible answers to the same navigational query.

One plausible idea is that if the two experts are, from your perspective, equally likely to answer the question correctly, you can't defer reasonably to either one. Maybe you can flip a coin, randomly choosing to defer to one air traffic controller. Doubtless, in view of your perilous circumstances, deference to one side would be prudential: You want to land the aircraft safely, and staying on your current flight path is not recommended by either expert. But a coin flip wouldn't permit you to believe that one expert is more likely than the other to be correct.

On the other hand, you may be positioned to defer reasonably to one expert if you break the symmetry between the two experts by having grounds to think they are not equally likely to get the right answer. That is what I call *asymmetry evidence*. Asymmetry evidence is an indicator, or grounds for believing, that one expert is more likely than the other to correctly answer a question. To illustrate, I will add a further detail to the original example. After you realize the air traffic controllers disagree, you remember that one control tower is around 50 miles closer to your present location. The remembered difference is asymmetry evidence. It suggests that the nearest controller should be trusted, given your background knowledge that radar systems become increasingly less reliable when detecting aircraft farther away. Now you can explain the conflict, and perhaps what you know makes deference to one of them reasonable for you.[6]

Importantly, having asymmetry evidence is only part of what you need. Asymmetry evidence is an indicator that one expert is more likely correct than the other. But any piece of asymmetry evidence may be neutralized by

The radio hoaxer was dubbed the "Roanoke Phantom." One of his messages, if a pilot had heeded it, would have led the aircraft to crash into nearby mountains.

[6] I should add that asymmetry evidence won't always explain the expert conflict. Suppose a trustworthy advisor tells you one expert is more likely to be correct than another expert. You are given no explanation for why that is so, but your asymmetry evidence could make deference reasonable for you. Even so, having some sort of explanation for the expert conflict is the customary type of asymmetry evidence and I'll focus on it here.

other pieces of evidence. So it must be the case that your first-order evidence, taken in its entirety, supports the view that one expert is more likely correct than the other.

To see why, recall your situation. You believe one air traffic controller is using a radar system located closer to you than the other controller, so the one nearby is more likely to give accurate navigational advice than the distant one. You have "location" asymmetry evidence. Let's assume you are correct about all of that—your evidence is accurate or non-misleading. But now imagine you realize that, for all you know, there could be an important asymmetry between the two controllers: The nearby controller is based in a small town, but the distant controller is based at an airport in a large city, where there's potentially a much more powerful radar system. You recognize that if the one radar system is in fact more powerful, then the distant expert may be more likely to provide accurate advice than the one nearby, or the two are equally likely to impart accurate advice. You have some "radar power" asymmetry evidence, which raises doubts in your mind about the significance of the "location" asymmetry evidence.

Here's what this means for you. Using the "location" asymmetry evidence as a basis for deference now depends on your being able to distinguish between two distinct states of affairs: (1) being nearby makes the nearby expert relatively more accurate than the distant expert and (2) the relative location of the experts does not ultimately change their relative accuracy. You are trying to get the most reliable navigational advice, and you rely on a proxy: evidence about the locations of the two air traffic controllers. At this point, your question is simple: Is the proxy I've chosen signal or noise? If you can't tell, then your initial asymmetry evidence does not settle the matter of which controller to trust.

You may be able to gather more evidence, of course. Perhaps you radio the two controllers again and learn they are using identical radar systems. In that event, if you can trust the controllers' reports about their respective radar systems, your "radar power" asymmetry evidence is eliminated and you have no reason to doubt that the "location" asymmetry evidence is accurate and non-misleading. Consider an account that expresses the idea that you can now defer reasonably:

> *Simple Account*: When you consult two conflicting experts, E1 and E2, concerning a question, you can defer reasonably to E1 if and only if (1) your

asymmetry evidence positions you to believe E1 is more likely than E2 to answer the question correctly.

The simple account is not correct. Even if condition (1) is necessary for reasonable deference, it is not sufficient.

To see this, consider the condition of hypoxia. Pilots in small aircraft flying at high altitudes may suffer from hypoxia when they become oxygen-deprived.[7] Hypoxia debilitates our reasoning without "leaving a trace" in our consciousness. Hypoxic pilots normally think they're perfectly sharp even when they are making grave errors. Let's assume you now have evidence to defer to the nearby air traffic controller and that your evidence is non-misleading. But then you suddenly recognize you may be hypoxic—one of the air traffic controllers mentions the possibility you are flying a little too high. You realize your assessment of the relative expertise of the controllers may be off-track. As a result, you come to have reason to think that hypoxia may have hampered your evaluation of your asymmetry evidence.

If you can't eliminate the credible doubt that you are hypoxic, you can't defer reasonably to one expert. That is true even though, as we are assuming, you are in fact correct to believe the nearby expert is the more likely one to deliver accurate navigational advice. Now let's imagine you activate an O_2 detector on board and learn that, thankfully, you aren't in low-oxygen conditions. What you've learned eliminates your doubt about your ability to evaluate your evidence properly.

Let's recap your situation. Initially, your reliability as a judge of your evidence was called into doubt because you believed hypoxia could be influencing your thinking. You addressed the doubt by gaining evidence to believe you are not hypoxic. This information allowed you to rely on your judgment concerning facts about how the experts' locations influenced their relative accuracy. You know you are good at making such judgments, we can assume, in virtue of your flying experience and training. So apparently you can now defer to the one expert over the other.

Two details should be underlined. First, you had a credible doubt about being hypoxic. But then you checked your O_2 detector and thereby gained reason to accept that you reliably judge whether the asymmetry evidence is accurate or non-misleading. So your "location" asymmetry evidence helped you defer reasonably to one controller. Second, there is a kind of normative

[7] Adam Elga (2008) introduced an example along these lines.

"trigger" that, when squeezed, demands higher-order evidence to affirm your reliability as a judge of the asymmetry evidence. Your doubt about hypoxia required you to get evidence of your judgmental reliability. You need not always seek out higher-order evidence of your judgmental reliability because your reliability is not always credibly called into doubt.[8]

At this point, we may expect you can finally defer reasonably to one controller. Here is a modified account of reasonable deference:

> *Reliability Account*: When you consult two conflicting experts, E1 and E2, concerning a question, you can defer reasonably to E1 if and only if (1) your asymmetry evidence positions you to believe E1 is more likely than E2 to answer the question correctly and (2) you have reason to believe you are a reliable judge of your asymmetry evidence if your reliability is credibly called into doubt.

But even if conditions (1) and (2) are individually necessary for reasonable deference, they are still not jointly sufficient. Consider one problem for the reliability account.

Suppose you find out the storm has produced atmospheric circumstances that degrade the accuracy of radar systems. For example, if you learn that 99% of radar systems are massively unreliable in this weather, you can't defer reasonably to one expert. That's true even if you know one expert is more likely than the other to deliver accurate information in normal weather. Perhaps the nearby expert is your best bet of the two, but if you defer to that one, your deference is still unreasonable—you should think that expert is likely unreliable in this weather.[9] Condition (1) requires that a novice believes one expert is relatively more likely to be correct than the other. We need a further condition stating that the expert you defer to does not, so far as you know, fall below some absolute threshold for being likely to be correct.

Here's a modified account of reasonable deference:

[8] I say that credible doubts call for higher-order evidence in favor of your reliability as a judge of the asymmetry evidence. An alternative requirement is that we must always gain higher-order evidence of reliability. That requirement threatens to induce widespread skepticism. For more on the issue, see David Christensen's helpful discussion of two types of belief-revision principles (2011, section 6).

[9] Let me emphasize that we care about epistemological evaluation here and that believing an expert's testimony can be subject to many types of evaluation. You may be morally, prudentially, or professionally permitted or required to believe something that is epistemically problematic.

Threshold Account: When you consult two conflicting experts, E1 and E2, concerning a question, you can defer reasonably to E1 if and only if (1) your asymmetry evidence positions you to believe E1 is more likely than E2 to answer the question correctly, (2) you have reason to believe you are a reliable judge of your asymmetry evidence if your reliability is credibly called into doubt, and (3) you do not have reason to believe that E1 is unlikely to answer the question correctly.

Doubtless, we could continue to refine this account of reasonable deference.[10] Questions about the account remain. For one, it treats reasonable deference as a categorical, all-or-nothing affair; but reasonable deference obviously admits of degrees. How should we think about that? How should we assign a particular level of confidence to our deference? For now, I say nothing more about the account. It's a serviceable conception of reasonable deference, and it will help us understand why the problem of conflicting expert testimony is so daunting.

Is Reasonable Deference Easy or Hard?

The "easy or hard" question concerns novices' ability to defer reasonably on the basis of their total evidence. The question is whether novices' actual evidence positions them to satisfy all the conditions for reasonable deference in situations where they would be confronted by expert conflicts. If novices'

[10] Let the refining continue here in this footnote. Consider a case that requires an extra condition for the Threshold Account. Suppose you are confronted by the conflict between the air traffic controllers, and again you have "location" asymmetry evidence: One radar system is closer to your present location than the other. On reflection, you have no reason to accept you are a reliable judge of the accuracy of your asymmetry evidence, but you have no credible doubts about your competence. Furthermore, you have no reason to think the nearby controller is unlikely to be correct.

So far, conditions (1)–(3) of the Threshold Account are met. But they seem to be insufficient for reasonable deference. Suppose you know you are oblivious to credible doubts about your own judgmental reliability. You recognize that even though open-minded, informed observers would have credible doubts about your reliability in this situation, you don't and can't. We can even imagine that you wouldn't come to doubt yourself even if such observers told you about their doubts. Your obliviousness is total. This means condition (2) is satisfied—not because you have reason to think you are reliable but because, as you know, you are insensitive or impervious to credible self-doubt. Plausibly, once you have reflected on your obliviousness to self-doubt, it won't be reasonable for you to defer to one expert on the basis of the asymmetry evidence. Instead, you should be unsure whether your asymmetry evidence is accurate or non-misleading or whether it's misleading but you just fail to recognize that.

I suggest accommodating this case by appending to the Threshold Account an extra necessary condition: You do not have reason to believe that you are insensitive to credible doubts concerning whether you are a reliable judge of your asymmetry evidence.

actual evidence would position them to defer reasonably to one or other of the rival experts a great deal of the time, reasonable deference is relatively "easy" for them. Alternatively, if they would often need to seek out more evidence and knowledge in order to defer reasonably, then doing so is relatively "hard." If reasonable deference is hard for some novices, different responses to expert disagreement will often be appropriate for them. (I discuss three alternatives below in "Lessons for Novices.")

My basic answer to the "easy or hard" question goes as follows. For novices who are informed and reflective, reasonable deference will still be hard in a great many situations. Reasonable deference demands work. Novices will need to expand their sets of evidence, adding evidence to believe they reliably judge the accuracy of their asymmetry evidence. The reason is that all novices are often at risk of falling into a situation where reasonable deference demands additional evidence of judgmental reliability; and informed, reflective novices often will—in virtue of what they know—find themselves in such a situation. Here's the upshot for all of us—decent candidates for being informed, reflective novices about some topics if there ever were. We will often need to get additional reliability evidence in order to defer reasonably.

To develop my case for that contention, I focus on condition (2) from the Threshold Account—namely, you have reason to believe you are a reliable judge of your asymmetry evidence if your reliability is credibly called into doubt. I call this the *reliability condition*. Why focus on this particular condition rather than the others? It's plausibly the most evidentially strenuous condition to meet, once its antecedent is satisfied. Comparatively, condition (1) will be easy to satisfy: Novices can generate asymmetry evidence on the fly. I also set to the side condition (3). If reasonable deference is like a chain, the reliability condition is the weak link.

Past work on the problem of conflicting expert testimony has focused almost exclusively on condition (1): that your asymmetry evidence positions you to believe one expert is more likely right than another. I call that the *asymmetry evidence condition*. In an article titled "Experts: Which Ones Should You Trust?" (2001), Alvin Goldman identifies five categories of empirical evidence that may sometimes position a novice to justifiably discriminate between rival experts. Goldman's five types of evidence include (1) arguments presented to a novice by experts to support the experts' own opinions; (2) the agreement from additional putative experts on one side of the question; (3) the evaluations of "meta-experts" concerning the experts' expertise, including experts' formal credentials; (4) evidence of the experts'

interests and biases concerning the question; and (5) evidence of the experts' track records.

How does this sort of evidence help a novice defer? To illustrate, return to the opening example involving anthropogenic climate warming. I described myself as having evidence of a scientific consensus. I also know about research by social scientists and historians, such as Naomi Oreskes and Erik Conway (2010), on the influence of oil-industry funding for scientists who deny the consensus view. The oil industry has shaped public perceptions of a "controversy" over climate warming by bankrolling the advocacy work of pundit scientists, demagogues, and empty suits. True, I am a climate-science novice. But I have empirical evidence about the experts, and this lets me defer reasonably to one side, assuming the other conditions for reasonable deference hold.[11]

Goldman's five types of empirical evidence can be thought of as rough-and-ready norms or principles, equipped with "other things being equal" clauses. One norm says that if you learn that two experts disagree and only one has financial incentives to accept a particular view, then, other things being equal, you have reason to think the other expert is more likely to be correct. Another norm says that if two experts disagree, and one has been correct about these matters much more often in the past than the other, then, other things being equal, you have reason to think the one with the better track record is more likely to be correct. And so forth. I mention norms for evaluating relative expertise only to observe that, in order to generate asymmetry evidence, the novice will ordinarily use such norms to draw inferences, implicitly or explicitly, from pieces of empirical evidence. Merely getting such evidence in hand will let the novice satisfy the asymmetry evidence condition.

In complex situations, a novice's total evidence concerning disputing experts will be settled by a subtle balancing act. One norm tilts toward this expert, another norm tilts toward that expert, and the novice's resting place in judgment depends on assessing the mixed body of evidence. There are difficult questions about whether our naïve norms deliver accurate judgments and about how good we are at assessing complex bodies of evidence, concerning putative expertise. Our norms are sometimes skewed. That's unsurprising because some norms make use of imperfect cognitive tools for

[11] Elizabeth Anderson (2011) defends a set of criteria for lay assessment of scientific testimony and uses the case of anthropogenic climate change as her main example.

perceiving bias. I'll say nothing more about the matter here, though the topic deserves attention.

Novices can display remarkable facility at generating putative asymmetry evidence, especially when they prefer one expert's viewpoint. Experts have noses of wax—novices tweak those noses as they wish.[12] In one of the opening examples, I described cancer patients and families who appealed to oncologists to prescribe mega-doses of vitamin C. These desperate people wanted to trust Linus Pauling instead of the medical establishment, and they invoked Pauling's impressive Nobel Prize. Who are you to disagree with a Nobel Laureate?[13] But Pauling misled the novices, and, plausibly, the novices easily could have known better. The reliability condition sets the bar for reasonable deference much higher than merely generating asymmetry evidence.

The reliability condition includes a kind of "trigger," as I noted. If novices' reliability as judges of their asymmetry evidence is called into doubt, then they need higher-order evidence to affirm their judgmental reliability. I use the term *reliability evidence* to refer to that higher-order evidence of competence. Once a novice's judgmental reliability has been credibly called into doubt, the novice can defer reasonably to one expert on the basis of the asymmetry evidence only if the novice gains sufficient reliability evidence to believe the following proposition:

> *R*: You are a reliable judge of the accuracy or non-misleadingness of the asymmetry evidence.

If the novice should disbelieve *R*, suspend judgment on *R*, or otherwise remain unsure about the appropriate attitude to hold toward *R* (Ballantyne 2019, pp. 109–115), the reliability condition is not satisfied.

There is the bad news for novices. They are often at risk of falling into a situation where they should doubt whether *R* is true. That's because it is fairly easy for them to gain evidence that challenges *R*, making doubts concerning *R* credible. But whenever they have credible doubts, they must gain reliability evidence in order to defer reasonably.

[12] As Alan of Lille, the 11th-century French theologian, wrote in his *A Defense of the Catholic Faith Against Heretics* of 1185–1200: "Now since authority has a nose made of wax—one that can be twisted in any direction—it needs to be strengthened with reasons." (Thanks to Peter King for the translation from the Latin.)

[13] Some scientists were overly deferential to Pauling. As J. D. Watson remarked, Pauling's fame made others "afraid to disagree with him. The only person he could freely talk to was his wife, who reinforced his ego, which isn't what you need in this life" (1993, 1813).

In the next three sections, I describe three types of evidence for doubting *R*: (1) facts about the tendency for novice assessments of expertise in a domain to be biased by novices' lack of knowledge, (2) facts about the tendency for novice assessments of expertise to be biased by novices' values, and (3) facts about the risk in some social circumstances for people to intentionally manipulate novices' norms and evidence in order to "manufacture deference."

Ignorant Novices

Work on the Dunning-Kruger effect provides evidence for doubting whether *R* is true. The Dunning-Kruger effect describes how ignorance delivers a "double curse": Our first-order ignorance tends to encourage second-order ignorance of our ignorance (Dunning et al., 2003; Kruger & Dunning, 1999). Across a surprisingly wide range of situations, people who perform poorly in a domain of knowledge tend to lack knowledge of their status as poor performers. The classic lesson from the Dunning-Kruger literature is that self-judgment is biased. In further work, David Dunning (2015; Dunning & Cone, 2022) has investigated how subjects' own knowledge influences judgment of other people's knowledge. If lacking knowledge leads to poor self-evaluation, how does it affect the evaluation of others?

Dunning and Cone discovered that subjects have "lopsided accuracy" in social judgment of expertise. Subjects more accurately evaluate the competence of people they outperform than people who outperform them. Knowing who knows less is easier than knowing who knows more. In Dunning and Cone's studies, average-performing subjects on knowledge-based tasks were better at correctly recognizing poor performers than top performers. Here's why. Subjects rely on their own knowledge[14] in order to evaluate the knowledge of other people. They tend to treat any deviation or departure from their own thinking as evidence of other people's incompetence. For example, when average subjects are assessing low performers, they interpret deviations from their own views as incompetence in the low performers; and since average subjects are assessing low performers, these judgments are basically right on track. But when average subjects instead judge top performers, the average subjects still treat deviations as evidence

[14] I use the term *knowledge* here to include mistaken and unreasonable beliefs.

of incompetence. That turns out to be a mistake: The fact that top performers deviate from average thinking tends to be a sign of special insight, not ineptitude. Consequently, low and average performers in some knowledge domains can't effectively distinguish top performers from the rest and often incorrectly rate average performers higher than top ones. As Dunning and Cone note, "genius, in our data at least, hid in plain sight. . . . For experts, it took one to know one" (ms, p. 17). There's empirical support for Jonathan Swift's quip, used as an epigraph for this chapter, that a true genius can be recognized by the confederacy of dunces who oppose them.

To give you some sense of the evidence supporting these claims, I'll describe one study. Dunning and Cone examined chess players' assessments of other chess players. The participants, recruited from college chess clubs or online, had US Chess Federation rankings of at least 700—a typical ranking for "scholastic" players or advanced beginners. They were first administered a multiple-choice test, asking them to "choose a move" in a chess game situation, either near the middle or the end of a game. Participants had to choose which of four alternatives was the best move. After completing their own test, participants graded five tests, putatively filled out by other participants, and had to indicate whether the target player was right or wrong in choosing each particular move. After grading each test, the participants had to indicate the likelihood, out of 100%, that they would win a game against the target, lose against the target player, or draw. What the experimenters discovered was that top chess performers were more severely misjudged than were the worst performers. As participant expertise increased, accurate assessment of the target increased. Perfect-score participants thought they had a 49% chance of defeating the top scoring target, whereas participants scoring zero judged their chances of beating the top target around 72%. Only high-scoring participants had the expertise necessary to correctly assess the challenge posed by the top target.

So far, I have noted empirical evidence that bears on novice perception of expertise. How does it create doubts concerning *R*, the proposition that you are a reliable judge of the accuracy of your asymmetry evidence? When novices form views about some experts' relative credibility, they may examine statements and arguments given by the experts. In fact, Goldman and others have suggested that novices can sometimes justifiably judge the "dialectical superiority" of one expert over the other. As Goldman notes, the dialectically superior expert may appear to novices to dish out more apparent rebuttals to the other expert's apparent counterarguments and to give quicker

responses to the other expert's counterarguments (2001, p. 95). Goldman says that novices who witness the experts' argumentative performances can infer that one expert has greater expertise.[15]

To see how Dunning and Cone's research bears on questions about reasonable deference, imagine the following situation. You are a novice sizing up rival experts. You come to believe one expert is more likely to be correct because it seems to you that their dialectical superiority over their opponent was revealed in a debate you watched. If you were to learn of psychological research showing how expertise can "hide in plain sight," then you would have some reason to doubt that your asymmetry evidence is non-misleading. Your reason for doubt is that novice-level knowledge often leads novices to inaccurate judgments of relative expertise. Learning the psychological research would give you some reason to disbelieve *R*, suspend judgment concerning it, or otherwise become unsure what to think about it (Ballantyne 2019, 109–115). After all, you are deciding to whom to defer but you are relying on a proxy: facts about apparent dialectical superiority. In light of what you know about novice perception of expertise, why believe the proxy you have chosen is signal rather than noise?

Partisan Novices

I've argued that novices' lack of knowledge can bias their evaluation of expertise. Their values can do the same. Evidence that values bias the evaluation of expertise comes from cultural cognition researchers. *Cultural cognition* is the tendency for people's values to influence their perceptions of policy, risk, and related empirical facts. Dan Kahan, a psychologist and legal scholar, has investigated with colleagues the *cultural cognition thesis*: the idea that people are disposed to believe that behavior they find respectable and honorable is socially beneficial and that behavior they find disrespectable and base is socially detrimental (Kahan & Braman, 2006). Cultural cognition researchers try to explain highly polarized social debates. On the one hand, it's plausible that many partisans in such debates typically form beliefs about policy and risk due to the operation of the same basic psychological mechanisms—biased assimilation, the affect heuristic, the availability heuristic, and so forth.

[15] For discussion of some norms guiding judgments of dialectical superiority, see Matheson (2005).

On the other hand, partisans have diametrically opposed and highly polarized perceptions of good policy and risk. How could that be? What explains sharp conflict between partisans, given that they tend to be outfitted with the same set of basic psychological mechanisms? According to the cultural cognition thesis, it's the interaction of values with psychological mechanisms that produces polarized opinions. Cultural cognition researchers explain conflict over topics such as gun control, capital punishment, and vaccinations by appealing to the ways that people's moral and political values function in processing policy-relevant information.

One interesting line of research in this paradigm focuses on politically motivated reasoning. How we process information is not isolated from our values, and our values move our opinions in predictable patterns. Sometimes, our views about policy-relevant issues become a badge of group membership, a way of signaling that we belong. As a result, people end up being selective in how they credit information in patterns that are consistent with their groups' views. That is just *motivated reasoning*: the tendency to assess factual claims in view of some goal that's independent of their correctness (Ditto & Lopez, 1992; Kunda, 1990). Politically motivated reasoning involves a goal that researchers call *identity protection*: "the formation of beliefs that maintain a person's status in [an] affinity group united by shared values" (Kahan, 2016, p. 3). Briefly put, politically motivated reasoning involves a person's crediting or discrediting new information in accord with the impact it will have on fitting their beliefs with the beliefs of people in an identity-defining group, not some truth-related norms.

Politically motivated reasoning can influence novice evaluations of expertise. People tend to trust experts whom they believe share their values and worldview, distrusting experts they perceive to hold different commitments. These patterns of trust and distrust can be explained by the mechanisms of politically motivated reasoning if people selectively credit or dismiss expert testimony in patterns that fit the values of their identity-defining group. And that's precisely what Kahan and his collaborators have observed (Kahan et al., 2011). In one study, subjects were presented with statements putatively from highly credentialed scientists. Subjects were asked to indicate how strongly they agreed or disagreed with the claim that each scientist was an expert on a risk or policy issue. The experimenters manipulated the positions the scientists held on cultural and political values. Subjects treated the experts as credible or not depending on whether the experts supported or contradicted conclusions that were favorable to the subjects' own values. In other words,

subjects sorted the experts as trustworthy or untrustworthy by taking cues from their group's values. Novices are highly attuned to information about experts' characters, but the information they pick up on does not necessarily track experts' reliability.

Here is a story about how values can influence novices' evaluations of experts. Physicist Hans Bethe was a pioneer in nuclear physics, a leader in the Manhattan Project, and a Nobel Prize winner. The celebrated physicist occasionally felt flummoxed when he tried to explain the benefits of nuclear power to opponents, many of whom were not trained in science. Bethe remarked that convincing them was like "carving a cubic foot out of a lake" (Walker, 2006, p. 21). Bethe argued that every energy system has risks; but the risks of nuclear power were manageable and nuclear power could actually deliver more energy with less environmental risk than the alternatives. One historian recounts a story Bethe told about speaking to an audience in Berkeley, California: "After [Bethe] had presented his position on the need for nuclear power, a woman in the audience stood up, turned her back on him, and shouted, 'Save the Earth!' The crowd reacted, he said, with 'thunderous applause' " (Walker, 2006, p. 21). Let's hear it for the antinuclear novices! Their values prevented them from seriously considering Bethe's claims. Intoxicated with solidarity and righteousness, they spurned the physicist.[16]

Evidence of how values influence novice assessment of expertise should lead novices to doubt *R*. If you learn that novices tend to evaluate conflicting experts in line with how well their positions fit with the values of their identity-defining group, you should think, How convenient! To generate asymmetry evidence, novices often attribute biases to one expert, but this sort of dialectical maneuver may just be politically motivated reasoning. That's not a reliable method for judging expertise—unless there happens to be some correlation between clusters of values and expert reliability. There are important questions here about how we could learn that values are in fact correlated with expert reliability and how values might themselves be a source of evidence, but for now my contention is simple. Learning about this psychological evidence should prompt novices to wonder whether they reliably judge the accuracy of their asymmetry evidence.[17]

[16] Thanks to Benjamin Wilson for sharing this story.

[17] For more discussion of the epistemological implications of cultural cognition research, see Greco (2021).

Toxic Epistemic Environments

I have argued that facts about novices—both about their lack of knowledge and about their values—can be evidence that leads us to doubt whether *R* (the proposition that we are reliable judges of the accuracy of some asymmetry evidence) is true. Facts about our social environments can also compromise our ability to evaluate expertise. We sometimes learn that people seek to "manufacture deference" or sow doubts in our minds, with the goal of nudging us toward one side of an expert debate.

Naïve norms for evaluating expertise are typically public. Since the norms can be recognized by observers, non-experts can sometimes learn to perform and self-present in conformity with the norms. Non-experts can appear to be trustworthy when they are far from it. Examples of "BS artists" abound. In the United States, there are a number of partisan political organizations devoted to training pundits in the art of appearing credible on television. At "pundit school" you learn to smile and interrupt your interlocutors effectively, to wear the right clothes or hip glasses, to dodge tricky questions (Parker, 2008). Pseudoscientists receive advanced degrees from unaccredited universities. Crank researchers publish bogus articles in predatory and vanity journals where there are virtually no editorial checks on quality.[18] Scientists get hired by industry to shill for pro-industry positions in media interviews and congressional hearings (Oreskes & Conway, 2010). Nothing is new under the sun. As I already noted, the problem of conflicting expert testimony goes back at least to Plato, who had encountered those teachers of rhetorical persuasion, the Sophists. In ancient Athens, the Sophists helped paying fools appear wise. Athenian novices faced obstacles in choosing between rival experts.

If non-experts masquerading as experts is not depressing enough, novices can also find themselves in situations where people fashion and distribute misleading evidence about genuine experts. One well-known example has been dubbed "Climategate." In 2009, an email server at the University of

[18] John Bohannon (2013) describes his "sting operation" to try to publish bogus articles in open-access journals. The articles all had fatal errors that any competent peer reviewer would spot easily. One article was putatively authored by a researcher named Ocorrafoo Cobange, a biologist at the Wassee Institute of Medicine in Asmara, Eritrea. Cobange's paper described the anticancer properties of a chemical extracted from lichen. Both Cobange and the Wassee Institute of Medicine were totally fictitious. Worse, the paper itself was a meaningless pastiche of technical jargon, "a scientific version of Mad Libs" (p. 62). None of this stopped the *Journal of Natural Pharmaceuticals* from accepting Cobange's article. Bohannon's sting was wildly successful, placing many sham articles in journals hosted by publishing conglomerates such as Elsevier and SAGE Publications.

East Anglia in England was hacked. Emails belonging to climate scientists were leaked by climate warming denialists. At first, many media outlets reported the emails had revealed, or at least suggested, that anthropogenic climate warming is a vast scientific conspiracy. But according to eight official investigations in the United Kingdom and the United States, there was no scientific misconduct or wrongdoing. Even so, many novices came to doubt the credibility of the scientific consensus about climate warming. Denialists had perpetrated a cunning smear campaign.

When I was wrapping up work on this chapter, "fake news" became a topic of public and academic discussion (Lazer et al., 2018). Prominent examples of fake news are written texts designed to spread misinformation, but some fabricated stories are circulated online merely in order to generate webpage traffic and advertising revenue. Fake news has also made the leap from text to video. Video-editing technologies allow purveyors of fake news to create videos of interviews that appear legitimate. A team of computer scientists developed a system that records video of someone talking and, in real time, modifies that person's facial expressions (Thies et al., 2016). Other new technologies can modify speech and audio in no less startling ways.

It doesn't take too much imagination to anticipate what is likely in store.[19] Climate warming denialists may create videos of climate scientists appearing to confess some "conspiracy" of science, and they'll then spread the videos on social media platforms. Climate warming advocates may get even by making videos of denialists appearing to admit, cynically, they are just in it for the money. As the quality of counterfeit video improves, novices and experts alike will have trouble telling the difference between real and fake footage. The power of images to influence our perceptions of experts' credibility should not be underestimated.

We should not believe everything we see. Indeed, if novices have reason to believe they are in toxic epistemic environments where some people seek to manufacture deference or to spread doubt about particular experts' credibility, novices may have reason to doubt whether *R* is true. For instance, if novices come to think their evaluations of some experts may easily depend on misleading evidence, they should doubt whether their asymmetry

[19] That is what I wrote in late 2016 when working on the first draft of this material. Several years later, as this chapter goes to press, I doubt any imagination is required. We have entered the era of "deepfakes" (a portmanteau of "deep learning" and "fake")—videos altered using artificial intelligence–based techniques that appear to show things that didn't happen.

evidence is accurate. Once this happens, novices need reliability evidence in order to defer reasonably.[20]

I've now described three types of evidence that novices can easily acquire and which, once acquired, should cause them to doubt whether *R* is true. Informed, reflective novices will find themselves in the following predicament. For a great many recognized conflicts between experts, if we can defer reasonably to one side, we will need reliability evidence that offers grounds to believe we are reliable judges of our favored asymmetry evidence. For us, reasonable deference will often be hard.

Lessons for Novices

For those of us trying to become more informed and reflective about some topic, what reliability evidence is there, and how can we get our hands on it? There is no general, one-size-fits-all advice. That's because evidence for doubting *R* can only be countered, eliminated, or ruled out by learning about the specifics of an expert dispute. Return to the aircraft example. Your "location" asymmetry evidence indicated that the nearby air traffic controller was more likely to impart accurate advice than the distant one. Then you had a doubt whether you were well positioned to evaluate your evidence. You realized you could be hypoxic. Using your O_2 detector, you cast aside your doubt that you were in low-oxygen conditions. This was enough to shore up your judgmental reliability, ensuring that you could defer reasonably to one expert on the basis of your asymmetry evidence.

One general lesson is that informed, reflective novices who satisfy the reliability condition will have done their homework. To get reliability evidence,

[20] Let me briefly compare what I have said about toxic epistemic environments to an example given by Gilbert Harman, who argued that someone's knowledge can be eliminated by the mere presence of misleading counterevidence in their social environment. In his "assassination" example (1973, pp. 143–144), Harman stipulates that you know that a politician has been assassinated on the basis of reading an early-edition newspaper that correctly reports the event. Later in the day, the early-edition papers are pulled from the shelves, and the state-controlled media begins reporting—falsely—that the politician is alive. All of this happens unbeknownst to you. According to Harman, your toxic epistemic environment eliminates your initial knowledge. You don't learn anything new, but since you could very easily hear the false reports, you now lack knowledge that the politician was assassinated. You lose your knowledge, Harman says, even if you don't even read a misleading newspaper or hear the false reports from a neighbor.

The idea I've deployed here is similar but even more plausible: If we are aware that we may easily be in a situation where misleading evidence concerning disagreeing experts circulates around us, then we have reason to doubt that our favored asymmetry evidence is accurate.

novices need to learn about experts' disputes, their methods, and their enterprise of making knowledge. That work can take considerable time and energy. And so a second general lesson is that, when life is too short and too busy for us to meet the reliability condition for at least some of the issues we care about, we should adopt alternative responses to expert disagreement.

Realistically, reasonable deference is practically impossible for the vast majority of controversial issues. What are the alternatives? I can think of three main options.

First, we can defer unreasonably to one expert, lacking any epistemic reason to favor that one over the other. This response amounts to some kind of "blind trust," an attitude that some philosophers say we must sometimes hold toward testifiers.[21] Second, we can use a method to aggregate the experts' conflicting judgments—at least if such a method is available to us—and reach a view that's distinct from expert judgment on both sides. Third, we can abstain from holding a view, refusing to take sides in the experts' conflict, choosing instead to mind our own intellectual business. There are different ways to abstain. If one expert believes a proposition and the other disbelieves it, we may suspend judgment about it. In more complicated disagreements, where different experts hold each of the three doxastic attitudes (belief, disbelief, and suspension of judgment), we may become unsure what attitude to adopt (Ballantyne 2019, 109–115).

As I have argued, the conditions for reasonable deference won't be met in many situations. An upshot is that we novices will often have to reconcile ourselves to one of these three alternatives, at least until we gain reliability evidence. The alternatives may not be as satisfying to us as reasonable deference would be. But they are often the best we can manage as we try to be informed and reflective novices.

In addressing the "easy or hard" question, I ignored uninformed, unreflective novices. But doesn't the account I defended have implications for them? One possibility is that it makes their reasonable deference too easy. Suppose a novice meets the conditions for reasonable deference set down by the Threshold Account and then systematically avoids new evidence, burying their head in the sand. Surely this novice is not reasonable. But doesn't my account imply that their deference is entirely reasonable?

[21] See, for example, Baker (1987) and Hardwig (1991). (Thanks to Johnny Brennan for telling me about Baker's article.)

Here are three observations about the objection. First, someone's inquiry can be properly or legitimately "closed" when they have reached the goal at which inquiry aims.[22] Suppose the head-in-sand novice has properly closed inquiry. On that assumption, we should not insist that the novice's opinion is unreasonable. Instead, the novice seems to be entirely within their epistemic rights to defer to one expert. If that's how we understand the case, though, it does not appear to threaten the account of reasonable deference. Second, we can grant that the novice's ostrich-esque policy gives them reasonable deference—so long as they don't recognize that they are evading new evidence. If they become aware of what they're doing, they should begin to doubt that they are well positioned to evaluate their asymmetry evidence. If they know what they're doing, they will be unreasonable to do it. Third, and most crucially, the head-in-sand policy is fundamentally defective. It is implausible that anyone should aim to defer reasonably by any means necessary. We should not always value reasonable deference more than gaining new information about expert disputes or reflecting on our asymmetry evidence or the like. Good epistemic policies will include being an informed and reflective thinker, being open to new evidence, trying to defer reasonably, and so on. A balance must be struck. Even if we grant that the head-in-sand novice has some positive epistemic status for their deferential belief, we can still epistemically evaluate them harshly.

A Problem for Experts

The problem of conflicting expert testimony challenges our social and intellectual commitments. It calls us to reflect on the manifold ways in which novices and experts relate to each other. As I have argued, novices who are informed and reflective must seek out evidence of their own reliability. That recommendation concerns what novices should do. But progress in addressing the problem should also consider what experts should do. I conclude by turning the spotlight from novices to experts, noting how they figure into solutions to the problem.

Some researchers study how novices react to experts' testimony. How can experts share their findings so that non-experts don't miss the message? Why is misinformation so resistant to correction? This increasingly important

[22] For discussion of what is required to properly close inquiry, see Kvanvig (2011) and Kelp (2014).

field of research goes under the banner of "science communication," but it encompasses questions about how experts in any truth-aiming field can communicate with outsiders effectively (Jamieson et al., 2017; Lewandowsky et al., 2012; Schwarz et al., 2016). Research on science communication sometimes examines questions about how novices reach accurate or inaccurate opinions on the basis of expert testimony. The basic model is that experts are attempting to insert accurate opinions into novices' heads—maybe gently or maybe out of frustration for the tenth time. How can experts get novices to accept correct views?

If reasonable deference is essential for the functioning of liberal institutions, it isn't enough for experts to always "insert" accurate opinions into novices' heads. Good reasons must somehow get in there, too. And so we should not overlook a slightly different question: How can experts help position novices to defer reasonably? Researchers could examine the matter. Novices who defer reasonably must navigate their way through a thicket of evidence, guided by their norms. But which norms are good ones? How can novices learn to use those norms effectively? How should experts testify to limited and ignorant novices who are nevertheless trying their best to defer reasonably?

Science communication researchers need not sort out these questions all alone. Philosophers could join in, too. Scientists could describe the cognitive and social factors that create and sustain the problem of conflicting expert testimony. Philosophers could describe good intellectual conduct for novices and testifying experts. They could together devise ideas to guide novices and experts, with the hoped-for outcome that novices receive guidance that helps them defer reasonably more often than they do now. Effective testimonial practices need novices and experts to play their roles well. They must collaborate in order to achieve the outcome of reasonable deference. I am suggesting that understanding how testimonial practices can be effective also calls for collaboration.

References

Anderson, E. (2011). Democracy, public policy, and lay assessments of scientific testimony. *Episteme*, *8*(2), 144–164.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Augustine of Hippo. (1947). The advantage of believing. In L. Schop (Ed.), *Writings of Saint Augustine* (Vol. 2. L. Meagher, Trans.). CIMA Publishing. (Original work published 391–392)

Baker, J. (1987). Trust and rationality. *Pacific Philosophical Quarterly*, *68*(1), 1–13.

Ballantyne, N. (2019). *Knowing Our Limits*. New York: Oxford University Press.

Berlin, I., & Jahanbegloo, R. (1992). *Conversations with Isaiah Berlin*. Phoenix Press.

Bohannon, J. (2013). Who's afraid of peer review? *Science*, *342*(6154), 60–65.

Christensen, D. (2011). Disagreement, question-begging and epistemic self-criticism. *Philosophers' Imprint*, *11*(6), 1–22.

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., Way, R., Jacobs, P., & Skuce, A. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters*, *8*(2), Article 024024.

Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R. L., Verheggen, B., Maibach, E. W., Carlton, J. S., Lewandowsky, S., Skuce, A. G., & Green, S. A. (2016). Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, *11*(4), Article 048002.

Delbanco, A. (2012). *College: What it was, is, and should be*. Princeton University Press.

Dunning, D. (2015). On identifying human capital: Flawed knowledge leads to faulty judgments of expertise by individuals and groups. In S. R. Thye & E. Lawler (Eds.), *Advances in group processes* (Vol. 32, pp. 149–176). Emerald.

Dunning, D., & Cone, J. (2022). "The Cassandra Quandary: How Flawed Expertise Prevents People from Recognizing Superior Performance among Their Peers." Department of Psychology, University of Michigan.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87.

Elga, A. *Lucky to be rational*. (2008). Department of Philosophy, Princeton University.

Goldman, A. (2001). Experts: Which ones should you trust? *Philosophy and Phenomenological Research*, *63*(1), 85–110.

Greco, D. (2021). Climate change and cultural cognition. In M. Budolfson, T. McPherson, & D. Plunkett (Eds.), *Philosophy and climate change* (pp. 178–200). Oxford University Press.

Hardwig, J. (1991). The role of trust in knowledge. *The Journal of Philosophy*, *88*(12), 693–708.

Hardy, J. (2010). Seeking the truth and taking care for common goods—Plato on expertise and recognizing experts. *Episteme*, *7*(1), 7–22.

Harman, G. (1973). *Thought*. Princeton University Press.

Jamieson, K. H., Kahan, D., & Scheufele, D. A. (Eds.). (2017). *The Oxford handbook of the science of science communication*. Oxford University Press.

Kahan, D. (2016). The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. In R. Scott & S. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences*. John Wiley & Sons.

Kahan, D. M., & Braman, D. (2006). Cultural cognition and public policy. *Yale Law & Policy Review*, *24*, 147–172.

Kahan, D., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, *14*(2), 147–174.

Kelp, C. (2014). Two for the knowledge goal of inquiry. *American Philosophical Quarterly*, *51*(3), 227–232.

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.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Kvanvig, J. L. (2011). Millar on the value of knowledge. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *85*(1), 83–99.

LaBarge, S. (1997). Socrates and the recognition of experts. *Aperion*, *30*(4), 51–62.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Matheson, D. (2005). Conflicting experts and dialectical performance: Adjudication heuristics for the layperson. *Argumentation*, *19*(2), 145–158.

Offit, P. A. (2013). *Do you believe in magic? Vitamins, supplements, and all things natural: A look behind the curtain*. HarperCollins.

Oreskes, N., & Conway, E. M. (2010). *Merchants of doubt*. Bloomsbury.

Parker, A. (2008, October 24). At pundit school, learning to smile and interrupt. *New York Times*. https://www.nytimes.com/2008/10/26/fashion/26pundit.html

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick and the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy*, *2*(1), 85–95.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-time face capture and reenactment of RGB videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2387–2395). Institute of Electrical and Electronics Engineers.

Walker, J. S. (2006). *Three Mile Island: A nuclear crisis in historical perspective*. University of California Press.

Watson, J. D. (1993). Succeeding in science: Some rules of thumb. *Science*, *261*(5129), 1812–1813.

Woods, S. (2014). Anthony Bourdain on writing, hangovers, and finding a calling. *Men's Journal*. Retrieved June 8, 2018, from https://www.mensjournal.com/features/anthony-bourdains-life-advice-20140919/

11

Against Strawsonian Epistemology

Testimony, Self-Knowledge, Promising, and Resolving

Hilary Kornblith

Peter Strawson's distinctive approach to the free will problem (1962/1997) viewed the key to understanding the nature of freedom and responsibility as lying in the reactive attitudes. Our tendency to feel resentment and gratitude, anger and forgiveness, and the like is, as Strawson would have it, ineliminible. More than that, Strawson held that, even if we could, somehow, dispense with such feelings, this would undermine worthwhile human relationships. Human freedom and responsibility are not rooted in any metaphysical fact, according to Strawson. They are, instead, rooted in our tendencies to feel these reactive attitudes, attitudes which are an essential part of a human life worth living.

In recent years, a number of philosophers have adapted this Strawsonian way of looking at things to various epistemological issues.[1] Thus, for example, Richard Moran (2001), Elizabeth Fricker (2006), Benjamin McMyler (2011), Edward Hinchman (2014), and Berislav Marušić (2015) have argued that the key to understanding the epistemology of testimony resides in the feeling of trust we may have in other human beings.[2] Once we understand the way in which trust underlies the epistemology of testimony, these philosophers claim, we see that paradigm cases of properly believing what others say should not be explained as a matter of having adequate evidence for one's belief. Testimonial belief, on this view, is grounded in a way which is different in kind from, for example, perceptual belief or belief based on inference.[3] We may, in some cases, believe what someone says on the basis of

[1] The term *Strawsonian epistemology* is used by Willaschek (2013) to denote quite a different sort of epistemological view, one which is rooted in Strawson's "Freedom and Resentment" in a way not directly related to the views discussed here.

[2] It is Marušić (2015) who points out the Strawsonian roots of this approach.

[3] The issue here, and in Strawsonian epistemology generally, as I see it, is not so much whether these beliefs are based on evidence but whether a proper understanding of the epistemology of these beliefs will see them as different in kind from both perceptual and inferential beliefs. I will thus use

Hilary Kornblith, *Against Strawsonian Epistemology* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press. © Oxford University Press 2022. DOI: 10.1093/oso/9780197636916.003.0012

evidence that the testifier is reliable, but when we do so, on this Strawsonian view, we have an attitude toward the testifier which objectifies them and is incompatible with genuine trust. Such an attitude is utterly foreign to healthy human relationships, as these philosophers would have it, and no part of our typical testimonial interactions.

The Strawsonian approach has been applied, as well, to understanding the epistemology of deliberation about what to believe. Richard Moran (2001) argues that the knowledge we have of what we believe, when the belief in question is a product of deliberation, is not a matter of having evidence that we have that very belief; it is different in kind from knowledge based on evidence. Knowledge of what we believe can, on Moran's view, be evidentially based; but in such cases we are estranged or alienated from our beliefs, just the opposite of the relationship we have to our beliefs when we deliberate. Moran emphasizes the Anscombian origins of this view (Anscombe, 1957), but there can be little doubt that it has deep affinities with Strawson's approach to the free will problem as well.

Finally, Berislav Marušić (2015), in what is one of the most creative and far-reaching applications of Strawsonian ideas, takes a similar approach to some of the beliefs we form when making promises or resolutions to behave in certain ways. As Marušić notes, we frequently resolve to undertake projects which we know will be quite difficult for us and which we will be severely tempted to abandon well before our goals are met. Similarly, we can and do sometimes promise others to undertake such projects. In these cases, Marušić argues that we may properly believe that we will do what we resolve or promise to do, but such beliefs, he argues, are not based on evidence. As with other Strawsonian epistemologists, Marušić argues that the attitude one has in forming beliefs based on evidence would undermine the possibility of promising and resolving in these cases, and it would undermine our status as agents as well. If we are to properly understand the epistemology of such beliefs, according to Marušić, we must see them as different in kind from evidence-based beliefs.

talk of "being evidence-based" as shorthand for this longer description both to make exposition more straightforward and because this is the way the Strawsonian epistemologists themselves describe the situation. Whether the right way to characterize the epistemology of perception and inference is in terms of evidence, or, instead, something else is orthogonal to the concerns of this chapter. I myself do not favor thinking about these issues in terms of evidence but believe, instead, that they are better approached in terms of the notion of reliability. Still, none of that will matter for the issues discussed here. I have discussed this issue further in Kornblith (2015).

I believe that this approach to all of these issues is fundamentally mistaken. While Strawsonian epistemologists see evidence-based approaches as deeply in conflict with pervasive and rewarding features of human relationships, I believe that the evidence-based approach is actually required for such relationships. But, more than that, I will argue that it is the evidence-based approach, and not the Strawsonian, which gets the epistemology right.

Strawson on Freedom

Strawson draws a distinction between two different sorts of attitudes we might take toward others and our interactions with them. He refers to these as the *objective attitude* and the *participant attitude*. When we take the objective attitude toward someone, we think about the person and our interactions with them in much the way that a contemporary social scientist might. We think about what the person might believe, say, or do in the light of the available evidence that bears on those matters, and we seek to form as accurate an opinion as possible. Taking the objective attitude involves treating the other person, as far as one's beliefs about that person go, as an object of theory; and we do the best we can to make sure that our theories about other people are true.

When we adopt the participant attitude, on the other hand, we do not distance ourselves from others in the way that the objective attitude requires. And it is precisely because we do not always, or even typically, distance ourselves from others that we are susceptible to the reactive attitudes. We do not merely note that someone has betrayed our trust; we resent it. We do not merely note that someone has extended themselves to us at great personal expense; we feel gratitude. And so on.

As Strawson remarks,

> To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided, though *this* gerundive is not peculiar to cases of objectivity of attitude. The objective attitude may be emotionally toned in many ways, but not in all ways: it may include repulsion or fear, it may include pity or even love, though not all kinds of love.

> But it cannot include the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships; it cannot include resentment, gratitude, forgiveness, anger, or the sort of love which two adults can sometimes be said to feel reciprocally, for each other. (1962/1997, pp. 126–127; emphasis in original)

As Strawson sees it, taking the participant attitude presupposes that the individual toward whom we take it is free and responsible. Although we certainly can and do avoid taking the participant attitude toward some individuals all of the time, and toward many individuals at some times, Strawson's view is that the participant attitude is something we cannot simply avoid taking tout court. We thus inevitably presuppose that there is such a thing as free action—indeed, that there is quite a bit of it—and that people are quite frequently responsible for their behavior.

Now there is a great deal about this view that one might call into question, but, at this point, I want to simply take it as given and see how well it may serve as a model for epistemological theorizing.

The Epistemology of Testimony

We frequently accept the word of others, taking what they say at face value and incorporating it into our bodies of belief. Competing accounts of the epistemology of testimony will not only offer differing views about the epistemically important features of these interactions; they will typically offer competing theoretical accounts of what the most central features of these interactions are.[4] For Strawsonian epistemologists, the central cases of testimonial knowledge involve testimony between friends and intimates, and what is important about communication between such parties is that it is mediated by way of trust.

Thus, consider Elizabeth Fricker's remarks about the character of knowledge by way of testimony:

> When I take another's word for it that P, I trust her in a way that makes my relation to her different from when I treat her expressed belief merely as

[4] For example, see Michaelian (2010) for an account and a system of classification, which is informed by psychological work on deception detection.

> defeasible evidence. One might say that I treat her as an end, not merely as a means. A fortiori this contrast holds, when through background information possessed by me, and not by her, I treat the fact of her utterances as a reliable natural sign of what is asserted. Moreover, as suggested earlier, it is plausible to see the function of testimony—its proper means of spreading knowledge—as being through the mechanism of trust in the teller, when her act is taken to be what it purports to be, an expression of knowledge, which offers to the hearer an entitlement to believe on the teller's say-so. T [Fricker's account of testimonial knowledge] holds only for the relatively narrow category I have described. But it is a category which reveals the nature of the speech act of telling, and of testifying more broadly, and enables us to discern and describe a crucial means of knowledge-spreading which is a true epistemic kind. (2006, p. 607)

On Fricker's account, as on all Strawsonian accounts, knowledge by way of testimony constitutes an epistemic kind when there is a certain personal—in her case, moral—relationship between speaker and hearer. It is in virtue of the trust between these parties that knowledge by way of testimony is possible, and this cannot be explained as a matter of the utterance being treated as evidence of the truth of what is uttered.

Others highlight the importance of the personal dimension in cases of testimony in a slightly different way. As Richard Moran puts it, the right way to view such cases does not involve, most fundamentally, believing a proposition; it involves believing a person (2005, p. 2). Similarly, Berislav Marušić (2015, Chapter 7, esp. section 7.1) tells us that what is involved is trusting a person, where this is not only different from but incompatible with believing what the person says on the basis of evidence of their trustworthiness. Similar points are made in Hinchman (2014) and McMyler (2011).

These accounts contrast with views on which testimonial knowledge is explained in ways which are of a piece with accounts of knowledge by way of perception or by way of inference: They are, on such views, similarly evidence-based. Thus, if my wife tells me that she'll be home late for dinner, I believe what she says because the fact that she said it is evidence of its truth. I have a great deal of evidence of my wife's veracity. Our long relationship together has provided overwhelming evidence that she would not say such a thing were it not true. And so, when I believe her, I believe her on the basis of this evidence.

The evidence-based account does not make one's relationship with the speaker irrelevant to what one should believe.[5] Apart from the fact that relationships of long standing give one a great deal of evidence about the trustworthiness of the speaker, there is another important fact about testimony from intimates which bears on one's evidence: Intimates have a great deal more to lose than mere acquaintances should they provide false testimony. If a used car salesperson tells you that a car you are interested in buying is in excellent condition, there is no harm to one's personal relationship should the claim turn out to be a blatant lie since there was no pre-existing personal relationship to harm. If one's friend, spouse, or lover tells a blatant lie, however, this can severely compromise one's relationship; and since everyone is well aware of this consequence of telling a lie, it gives such individuals still greater incentive, over and above their background commitment to honesty, to be honest in such personal communications. Evidence-based accounts of testimonial knowledge thus do not ignore the epistemological relevance of personal relationships. They merely treat facts about such relationships as further pieces of evidence.

It is precisely this way of dealing with features of personal relationships that Strawsonian epistemologists object to. Marušić is admirably clear on this point:

> A simple objection to [evidence-based accounts][6] is that having adequate evidence to believe precludes the need for trust. In particular, if you have adequate evidence to believe that someone will do something, then there is no need to trust her to do it. Hence [an evidence-based account] fails to give an account of trust altogether; it misses the phenomenon it is supposed to explain. (2015, p. 180)

I would put this point somewhat differently. Those who favor an evidence-based account of testimony do not, of course, deny that there is such a phenomenon as trusting a person. Rather, they see the sort of trust at issue in testimony as nothing more nor less than evidence of trustworthiness or reliability. Thus, we may trust our friends and intimates just because we have such

[5] Marušić surely overstates his case on this matter when he remarks, "Since [an evidence-based account] fails to attribute any significance to the interpersonal relationship that is involved in trust, it should be rejected" (2015, p. 182).

[6] Marušić speaks here of "the Evidentialist Response," but I prefer to put this more neutrally, for reasons given above in note 3.

overwhelming evidence that they are worthy of trust and, therefore, that what they tell us can be relied upon to be true. Of course, the Strawsonian thinks that there is a different phenomenon going on here, and those who favor evidence-based accounts of testimony will, indeed, deny that there is such a phenomenon. But it is important to see this as a theoretical dispute about just what is going on in cases of testimony between friends and intimates.

Why does Marušić believe that evidence-based accounts fail to appreciate the nature of communication between intimates? It is here that Marušić draws on distinctively Strawsonian resources:

> there are two ways to answer the question of what someone else will do. We can answer the question as observers. We will then assess our evidence about what the other will do in light of the fact that she is making a commitment and seek to predict what she will do. Yet we can also, to draw on Strawson's notion, take a participant point of view. We can *ask* the other what she will do. And, if she promises us to do something, or tells us that she will do something, we can, at least in the good case, take her at her word and trust her. When we take an observer's view of the other's future, our belief is rational in light of our evidence; when we take a participant view of the other's future, our belief is rational in light of the reasons of trust. (2015, p. 192; emphasis in original)

Moreover, as Marušić emphasizes, "reasons of trust" are "categorically different from evidence" (2015, p. 183).

Now I do think that Marušić has got the phenomenology of communication between intimates exactly right. When my wife tells me that she's going to be late for dinner, I don't stop to think about the extent of her trustworthiness or the amount of evidence I have from her behavior over a period of many decades that she will tell me the truth. Nor do I think about the ways in which our relationship would be harmed were she to lie to me and the fact that she is aware of the harm that such violations of trust can cause in a relationship. None of these things cross my mind. I just trust her.

But, of course, none of these facts about the phenomenology of trust have anything to do with the question of what my testimonially based belief is ultimately based on.[7] Thus, consider the fact that we not only have friends and intimates who are eminently trustworthy but many of us have personal

[7] For a useful discussion of these issues, see Miranda Fricker (2007, Chapter 3).

relationships with individuals who are less than fully trustworthy. Indeed, rather than seeing trustworthiness as a simple yes/no matter, it is no doubt more accurate to see the matter of trustworthiness as spanning a wide range of cases, from individuals who are exceptionally reliable to individuals who are exceptionally unreliable. Most people lie somewhere in between these extremes, even when one factors in, as one must here, the context of communication with an intimate. It would be foolhardy not to be sensitive to these differences and not to take them into account in one's epistemic interactions with others. And, of course, we do take these things into account.[8] In cases where our relationship with the speaker is close and of long standing, these matters can be taken into account without having to bring them to conscious attention. I don't need to think about my wife's trustworthiness or her commitment to be honest with me in order for this to play a role in my acceptance of what she says, any more than I need to focus attention on the untrustworthiness of the snake-oil salesperson in order to be uninfluenced by their mendacity. Indeed, we can only account for the differential way in which the testimony of others affects us by recognizing the role which these background beliefs play without having to be brought to mind.

Those who favor an evidence-based approach to testimony will thus see the proper role for the recipient of testimony as residing in the objective point of view. We should simply be responsive to the evidence available and form the appropriate conclusions on its basis. Evidence about our relationship to the speaker is relevant, of course, in just the ways I have enumerated. But it is the objective attitude which rightly describes the way in which we should respond to testimony.

Now if we accept some of Strawson's claims about the consequences of viewing others from an objective, rather than a participant, point of view, this will certainly raise various practical concerns. If Strawson is right, for example, that the objective point of view undermines satisfying human relationships, then if evidence-based accounts of testimony are correct, we seem forced to choose between being epistemically responsible in dealing with the testimony of intimates, on the one hand, and having satisfying human relationships, on the other. Marušić encourages this view when he says that when we think of others from the objective point of view, we "treat the other as an object" (2015, p. 200). Similarly, he quotes Moran approvingly

[8] Indeed, we begin taking such things into account as very young children. See Harris (2012, especially Chapters 5 and 6).

when Moran remarks that "refusing to acknowledge an epistemic stance toward the speaker's words other than as evidence means that speaker and audience must always be in disharmony with each other" (Moran, 2005, p. 23; quoted in Marušić, 2015, p. 182).

But this way of thinking about the matter surely provides us with a false dichotomy. Consider my response to my wife's telling me that she will be late for dinner. Taking the objective stance does not require treating her as an object. For one thing, it is just an objective fact that my wife has mental states, so the objective stance will certainly involve thinking of her as a person. And while there would certainly be something very odd, and alienating, if I were to self-consciously rehearse the reasons for thinking that my wife does not lie to me,[9] this too is completely irrelevant to whether I am adopting an objective attitude toward her. If my belief that she will be home late is based on evidence in the right sort of way, then I am adopting the objective attitude. I do not see any reason to think that someone who is moved by the evidence in this sort of way would have their relationship with the speaker thereby compromised. In cases where the relationship with the speaker is a good one, the kind of etiology for belief that an objective attitude requires simply does not threaten to produce the kind of alienation and distance which Strawson, and Strawsonians, suggest.

One might think that alienation threatens, with its attendant harm to personal relationships, in cases where an objective attitude would provide less confidence in the speaker's assertion. There is, of course, a range of cases here. Sufficient acquaintance with the speaker will not require that the hearer self-consciously rehearse any reasons in order to take them into account, so that cannot be the worry, wherever the case may fall on the spectrum from extremely trustworthy to extremely untrustworthy. One may certainly have close relationships with people who are highly trustworthy in general and yet are not to be trusted on certain topics. If my friend Jerry, who is otherwise highly trustworthy, has a huge blind spot about his children, always viewing them in the best possible light and never seeing their shortcomings, however obvious those shortcomings may be to others, then I may simply discount his joyous remarks about his children's latest accomplishments, without having

[9] The worry that the objective attitude would require such a conscious rehearsing of reasons is reminiscent of Bernard Williams's "one thought too many" argument against utilitarianism (Williams, 1981, Chapter 1). (Indeed, there is some reason to think that Marušić may have Williams's argument in mind since he alludes to the question of what to do when one's spouse is drowning, the very example which is the focus of Williams's discussion [Marušić, 2015, p. 199].) For an illuminating discussion of the "one thought too many" argument, see Baron (1984, especially pp. 211–214).

to mentally rehearse the reasons for failing to take them at face value. But this needn't get in the way of having a meaningful relationship with Jerry. If such a thing did stand in the way of meaningful and satisfying relationships, then the world would present very few opportunities for such relationships. Most people find little trouble negotiating such blind spots and minor failings, both moral and epistemic, in their friends and intimates.

Of course, Strawson is right that, in extreme cases, an objective attitude may require the kind of emotional distancing which makes a close personal relationship impossible. There are all sorts of character disorders which, once one recognizes that a speaker is subject to them, will make it impossible to have a close personal relationship with that person. But here it is not the objective attitude but rather the character disorder itself which is the source of the problem.

Strawson's worry, then, that objectivity of attitude will inevitably lead to regarding others as mere "object[s] of social policy," and thereby undermine the possibility of satisfying human relationships, is badly misplaced. By the same token, those who would leverage Strawson's ideas about the reactive attitudes into an account of the epistemology of testimony have a terribly inaccurate idea of what being responsive to evidence requires and reject an evidence-based account of testimonial knowledge on the basis of that misunderstanding.

The Epistemology of Self-Knowledge

Richard Moran (2001) has applied Strawsonian ideas to the epistemology of self-knowledge. Moran is interested, in particular, in the phenomenon of deliberation, especially about what to believe, as well as the knowledge we have of what we believe as a result of the deliberative process. As in the case of Strawsonian views about testimonial knowledge, Moran is concerned that viewing one's beliefs about what one believes as evidentially based will result in a certain sort of alienation—or estrangement, as he puts it—in this case, from oneself, as opposed to the alienation from others which Strawsonians worry about in the case of testimony.

Consider Gareth Evans's (1982) famous example. If you ask me whether I think there will be a third world war, I may stop to consider the issue. When I deliberate here, I think not about my mental state, what it is I believe or think; instead, I think about the world's political situation. Here, we have the

first-person perspective of the deliberator, the Strawsonian participant perspective. I might, however, direct my attention to my mental states, as well as the behavior which provides evidence about them. Here, I treat myself in much the same way that I would treat you if I were asked what you believe on a certain matter. Here, I adopt the Strawsonian objective perspective. Moran insists that the deliberator's perspective on their own beliefs cannot be evidentially based: "It is part of the ordinary first-person point of view on one's psychological life . . . that evidence is not consulted" (Moran, 2001, p. 92).

Just as in the case of testimonial knowledge, it seems clear that the Strawsonian account certainly has the phenomenology right. There is all the difference in the world between self-consciously focusing on evidence about what my mental state is—for example, whether I have good evidence that I believe there will be a third world war—and what goes through my conscious mind when I deliberate about whether there will be such a war, where my conscious attention is focused on evidence about the political scene. In general, if you ask me what I believe about various issues, I am not consciously aware of attending to evidence about my mental state.

And just as with the case of testimonial knowledge, it should be clear that what I am consciously aware of tells us very little about what it is that my belief is actually based upon. Indeed, a great deal of evidence has been offered by a variety of psychologists (see, e.g., Gopnik, 1993, for discussion; also see Carruthers, 2011; Cassam, 2015) that our beliefs about our own beliefs, and about our propositional attitudes generally, are, indeed, evidentially based. This view is not uncontroversial, but there can be little doubt that if the view is to be rejected, something more than an appeal to how things seem from the first-person perspective will be required.

But much as Moran does seem simply to take for granted that the first-person perspective on our mental states is not evidence-based, it would be unfair to leave it at that. There is a contrast which Moran is getting at which deserves further exploration and which does not presuppose anything about the basis of judgments about our own mental states that are a product of the first-person perspective.

There is an important distinction between avowing a belief and reporting a belief. One might, as a result of extensive psychotherapy, come to recognize that one believes that one's brother, who disappeared long ago, is dead. That one has this belief is very clearly the best explanation of one's behavior; alternative explanations all face utterly conclusive refutations. In such a situation, one would be able to report what it is that one believes. But this does not

assure that one would be able to avow it. One might, after all, recognize that one has a variety of irrational beliefs; and, in these cases, one could report that one had them—acknowledging, for example, that one should probably try to do something to overcome one's own irrationality. In such cases, precisely because one recognized these beliefs to be irrational, one would not be able to avow them. And if the question of the rationality of a belief remains unsettled, a similar conclusion might apply: Reporting that one has the belief would be possible; avowing it would not.

Now we will all need to be able to make this distinction, even if we think that the distinction should not be explained as due to a difference between beliefs which are based on evidence and those which are not. And there can be little doubt that the person who can report on the content of their belief but cannot avow it is properly described as somehow alienated or estranged from their own beliefs. But once we give up Moran's suggestion that the difference between the person who is estranged from their belief and the person who is not is due to the presence or absence of an evidential basis for that person's belief about their own belief, then what remains of the epistemological significance of the distinction between beliefs we can avow and those we can only report?

Consider the following passage from Moran:

> it does seem appropriate to distinguish between different levels at which one conceives of oneself as a psychological subject. Believing involves taking something as true; and of course, one also takes other people to have true beliefs sometimes. But the beliefs of other people represent facts (psychological facts, to be sure) on the basis of which one may make up one's mind about some matter, whereas one's own beliefs just are the extent to which one's mind *is* (already) made up. That is, the beliefs of another person may represent indicators of the truth, evidence from which I may infer some conclusion about the matter. I may trust or mistrust them. With respect to my own beliefs, on the other hand, there is no distance between them and how the facts present themselves to me, and hence no going from one to the other. (2001, p. 75; emphasis in original)

Moran is quite clearly talking here about beliefs one might avow since beliefs one is able to report on without being able to avow them are not examples of "how the facts present themselves"; they are, instead, psychological facts about oneself, epistemologically on a par with the beliefs of others.

Now, of course, one may avow a certain belief and then find oneself in a position where evidence is presented which undermines one's confidence in it. Someone might avow that a particular job candidate is not well suited for a given job and yet, on being presented with evidence about implicit bias, come to question whether the belief about the candidate should be maintained (Brownstein & Saul, 2016). Or one may avow a belief about how the check at a restaurant may be fairly divided and come to question that belief as a result of disagreement from others (Christensen, 2007). Or one may avow a belief as a result of deliberation about what to believe and entertain serious doubts about the matter as a result of learning about confabulation (Kahneman, 2011; Kornblith, 1999, 2012; Wilson, 2004). Beliefs which one avows are not immune to doubt, and when one is presented with reasons for doubt, one may come to look upon them as psychological facts about oneself which may or may not be good indicators of how things stand in the world outside one's mind. In situations of this sort—situations which are in no way exceptional—one finds oneself switching back and forth between being able to avow one's belief and only being able to report it.

This ability to shift perspectives on one's beliefs—to move back and forth between a Strawsonian participant perspective and an objective perspective—is absolutely crucial for any circumspect believer. It may be, as Moran suggests, that the individual who avows a belief regards the truth of that belief as settled (2001, p. 77); but even if this is the right way to think about avowal, regarding a certain matter as settled at a particular instant must then be fully compatible with regarding it as unsettled just a moment later. And if one regards individuals who take an objective perspective on their own beliefs as thereby alienated or estranged from their beliefs, then this kind of alienation or estrangement is an altogether healthy thing epistemically. Psychological evidence about the illusions which the participant perspective brings with it (Kornblith, 2012, 2013) can serve as an important corrective to the first-person perspective. Refusing to take such an objective perspective on one's own beliefs would be epistemically disastrous.

Perhaps Moran would not disagree.[10] Moran is concerned to distinguish between two different sorts of authority which we might claim to have over

[10] It's quite hard to say just what Moran's position is here. He quotes Wittgenstein's remark that "One can mistrust one's own senses, but not one's own belief" with approval (2001, p. 75). And he goes on to say, on the same page, that "this must mean not that I take my beliefs to be so much more trustworthy than my senses, but that neither trust nor mistrust has any application here." But the familiar points made in the previous paragraph of the text certainly suggest that raising questions about the trustworthiness of one's own beliefs not only makes perfectly good sense—and thus that it

our beliefs: epistemic authority and the authority of rational agency (2001, e.g., pp. 92–93). Claims about, for example, invulnerability to error or to certain types of error fall under the heading of epistemic authority; but what Moran calls the authority of agency involves the ability to avow a belief, and this ability does not flow from any degree of epistemic authority, no matter how great. One might have superb evidence that one has a certain belief and yet still be unable to avow it. What the Strawsonian participant perspective on one's belief allows for—the ability to avow the belief—is the recognition that one is the author of one's beliefs, that what we believe is up to us. As Moran sees it, when we take the objective perspective on our beliefs rather than the participant perspective, we thereby lose the authority of agency; and, with it, we fail to view ourselves as responsible for our beliefs.

When we deliberate about what to believe, as Moran sees it, we view what we will believe as up to us. We are the authors of our own beliefs, and the process of deliberation, on this view, is the method by which we take our beliefs in hand and take responsibility for them. Someone who takes the objective perspective on their beliefs loses out on this.

> For if he cannot see the empirical question of what he believes as answerable to his current explicit thinking about the matter, then just being informed what his belief is leaves open the question whether this information shall count as a *reason* for him or not. As it is, he is no better off than if he had been told that some other person has this belief, or that he himself once did at some other time. (Moran, 2001, p. 123; emphasis in original)

Let us consider, first, an individual who occupies the Strawsonian participant perspective on their own beliefs. Let us call him Sal. Sal deliberates about

would be wrong to say it has no application—but is an important part of an epistemically healthy life. In a footnote, Moran remarks,

> The ancient contrast between the seductive, misleading Senses and the trustworthy dictates of Reason can be seen, in part, as resting on a failure to recognize a related difference in kind between the two. The Senses can be compared to an unruly mob, in conflict with itself, because they belong to the category of deliverances *on the basis of which* one forms a judgment. But, insofar as Reason represents the unifying judgment one forms *from* this basis, it is not a faculty superior to or in competition with the Senses. (2001, pp. 75–76, n6; emphasis in original)

Moran seems to take this as reason for thinking that "from the first-person point of view, the relation between one's own belief and the fact believed is not evidential or empirical" (2001, p. 76) The idea that the reliability of reason cannot be treated as an empirical matter is similarly defended by Thomas Nagel (1997). I have discussed this matter in some detail in Kornblith (1999), where I argue that the reliability of reason can and should be viewed as an empirical matter.

what to believe on a certain matter. He evaluates the reasons for and against a certain proposition *p*, and, after reviewing the reasons carefully, he concludes that the evidence he has strongly favors *p* over *not-p*; he therefore comes to believe that *p*. Sal views himself as the author of his belief, and he views the reasons he rehearsed as his reasons for believing that *p* and, indeed, as good reasons to believe that *p*. Sal takes responsibility for his belief; it is fully his own; and, as is typical in the case of such deliberation, Sal will avow that *p* when asked what he believes on this matter.

Consider, now, a second individual, Cal, who also deliberates about whether *p* and goes through the same private monologue as Sal. She too concludes that *p*; she takes this belief to be based on the reasons which she rehearsed in her private monologue; and she takes these to be good reasons. But then she stops and thinks again. Cal remembers what she learned about the psychology of deliberation, and she takes an objective perspective on her belief that *p*, as well as on her deliberative process. She reminds herself that the process of deliberation does not very accurately track the beliefs on which one's conclusion is based. One's belief is based on a great many other beliefs which one did not bring to consciousness, and many of the beliefs which one did bring to consciousness may have had little effect on what one ultimately believes (see, e.g., Kunda, 1999, p. 308). In taking this perspective on her belief that *p*, Cal no longer regards the beliefs she rehearsed in the process of her deliberation as an accurate accounting of her reasons for believing that *p*, and she begins to wonder whether the reasons for which she actually believes that *p* are genuinely good ones. For this reason, Cal is no longer in a position to avow her belief that *p*.

Cal may consider the matter further and consult with various authorities on the psychology of deliberation. She may participate in a variety of experiments designed to determine even idiosyncratic features of her own psychology. Let us suppose that this is done in extraordinary detail and with great accuracy. Cal comes to understand, from the objective perspective, just what went on in her mind when she deliberated about whether *p*, and she comes to recognize exactly what her reasons were for believing it. These reasons are not ones which she can recognize from the first-person participant perspective, and Cal does not regard the reasons which were part of her private monologue when she deliberated as the ones which settled for her the question of whether to believe that *p* or *not-p*. Given what she knows about the deliberative process, she cannot take her private monologue under conditions of deliberation as accurately revealing what her reasons for belief

actually were. As Cal now sees things, the participant perspective distorts her view of her reasons. It is only the objective perspective which allows her to appreciate what her reasons were, even if, when she takes that perspective, she can only report, and not avow, that these were her reasons. Cal knows too much about the process of deliberation to take it at face value in the way that the participant perspective demands, and it is for precisely this reason that Cal cannot take her explicit thinking when she deliberated to be what settled the matter of her belief about whether *p*.

Does taking an objective view of one's processes of belief acquisition thereby undermine one's rational agency, prevent one from taking responsibility for one's belief, or limit one to reporting, rather than avowing, one's belief? It might. If an objective view of the way in which I arrived at a belief reveals it to be the product of non-rational or irrational processes, this may well be the result. Thus, in Nisbett and Wilson's (1977) classic paper on confabulation, a large percentage of subjects came to believe that a certain pair of pantyhose were the best of those presented, not, as they thought, because they had noticed certain good-making features of the favored pair but as a product of a tendency to favor objects on the right. Learning this about one's belief would certainly undermine rational agency; it would surely undermine any sense of responsibility for one's belief; and it would prevent one from avowing the belief, even if the objective perspective revealed that one continued to hold it.

Needless to say, not all beliefs are produced by non-rational or irrational processes, and so an objective perspective on one's processes of belief acquisition may reveal that a particular belief one holds was, in fact, rationally acquired. In cases of this sort, the objective perspective would in no way undermine one's rational agency, nor would it prevent one from taking responsibility for holding the belief. The discovery that one's belief was rationally produced would not in any way interfere with one's ability to avow it. Just as the objective perspective on one's interlocutor needn't get in the way of satisfying personal relationships, an objective perspective on one's beliefs and their etiology needn't get in the way of a view of oneself as a rational agent, responsible for one's beliefs.

Moran's Strawsonian worries, then, that the objective perspective on one's beliefs must inevitably result in alienation or estrangement from them are misplaced. Moreover, when the objective perspective does result in such alienation, because the manner in which one's belief was produced was either irrational or non-rational, the resulting alienation is, epistemically, a good

thing. It can be an essential first step to putting one's epistemic house in good order.[11]

The Epistemology of Promises and Resolutions

Berislav Marušić (2015) takes a Strawsonian approach to the epistemology of promises and resolutions. Marušić points out that people frequently make promises and resolutions to do things in cases where they have a good deal of evidence that they might not follow through. Thus, for example, a heavy smoker may resolve to quit smoking, even knowing that they have tried to quit several times, on each occasion going back to their old habits fairly quickly. Such private resolutions are surely not uncommon. By the same token, one may promise a loved one, in the very same situation, that one will quit smoking. Marušić is interested in whether one should believe, in these sorts of situations, that one will actually follow through on one's resolution or promise.

Following Marušić, let us stipulate that we are dealing with a case in which, if one were to take an objective attitude toward one's behavior, one could not reasonably believe that one will follow through. It would be epistemically unjustified to believe, in the kind of case Marušić has in mind, that one will actually quit smoking. This is not to say that one should believe that one won't. Rather, the evidence simply does not permit an epistemically justified belief that one will succeed in quitting.

Nevertheless, Marušić wishes to argue that it is permissible to believe that one will succeed in one's undertaking in this sort of situation, and the form of the argument he presents is straightforwardly Strawsonian. On Marušić's view, making a promise or a resolution requires that one take a participant attitude toward one's situation, rather than an objective

[11] Similar considerations apply to concerns about alienation in cases of decision-making. Thus, Paul (2014) argues, in Strawsonian fashion, that it would be a mistake to make important decisions, such as whom to marry or whether to have a child, by taking an objective perspective on one's decision. Paul insists that we must make our choice from the participant perspective:

> If we don't choose this way, then in an important sense, we alienate ourselves from our choices, and thus alienate ourselves from our own futures. In other words, if you don't make choices about your future from your own personal point of view, and instead attempt to map out choices based only on some sort of impartial, uncommitted, third personal point of view, you in effect cede authority over yourself. (2014, p. 130)

Just as in the epistemic case, this kind of distancing of oneself from the participant perspective seems to me, instead, to be the better part of wisdom.

attitude. Although an objective attitude would reveal that one should not believe that one will follow through, promising and resolving preclude taking an objective attitude toward oneself, and the participant perspective leaves room for beliefs which the objective perspective would rule out. As Marušić explains,

> Our view of what we will do, when matters are up to us, is made rational by our practical reasons; in contrast, our view of what will happen, when matters are not up to us, is made rational by our evidence. (2015, p. 123)

It will help to consider an example in some detail. Suppose that Ibrahim, concerned for Mary's health, asks her to promise him that she will quit smoking. It's not that Mary does not want to quit smoking; she very much does, both for her own health and out of respect for her relationship with Ibrahim. But in the past, this has not led to any progress in giving up the habit, which both she and Ibrahim view as a bad thing, all things considered.

Now if Mary considers whether she is likely to succeed in quitting this time, should she promise Ibrahim to quit, she can take an objective perspective on her behavior and her past performance. If she does this, she will find that her evidence does not permit her to believe that she will succeed in quitting. But taking this perspective, as Marušić sees it, would have Mary viewing her behavior as something over which she has no choice; it is simply dictated by her psychological make-up together with various facts about her situation. Taking the objective perspective thus amounts to viewing her own smoking as something which is not up to her; it thus inevitably involves ducking responsibility for her own behavior. In order for Mary to take responsibility for her smoking, she needs to adopt the participant perspective.

If Mary takes the participant perspective in response to Ibrahim's request, she will not be focused on features of her psychology or the ways in which they will causally interact with her situation. Instead, she will consider what the reasons are for quitting, and here, to be sure, all the reasons count in its favor. But this just means that Mary has reason to promise Ibrahim that she will quit. To take the participant perspective involves viewing one's behavior as something over which one has a choice, and when Mary views quitting in this way, and she considers the reasons she has to quit, "she can settle the question of what she will do by considering her practical reasons" (2015, p. 124). And what this means is that Mary should believe that she will quit.

The objective perspective and the participant perspective would lead one to different, and incompatible, beliefs in this situation. And it is for this reason that Marušić insists that one must simply resist taking on the objective perspective. Here, he quotes Bas van Fraassen with approval:

> I say, "I promise you a horse," and you ask, "And what are the chances you'll get me one?". I say, "I am starting a diet today," and you ask, "And how likely is it that you won't overeat tomorrow?". In both cases, the *first* reply I must give is "You heard me!". To express anything but a full commitment to stand behind my promises and intentions is to undermine my own avowals as a person of integrity and, hence, my entire activity of avowal. (van Fraassen, 1984, p. 254; quoted in Marušić, 2015, p. 152; emphasis in original)

Indeed, Marušić goes further than van Fraassen. He insists that "we should refuse to take a theoretical view of ourselves" (2015, p. 154).

Now, I admit that when someone says, "How do I know that you will do it?" in response to a promise, this can be rather rude. Still, there will be situations in which it is perfectly understandable that someone should ask this question; and, indeed, the situations Marušić has in mind, where there is a good deal of reason to doubt that the promissor will follow through on the promise, are just such situations. Since Mary has tried to quit smoking many times, with no success and no progress toward success, Ibrahim might understandably wonder what reason there is to believe that Mary will do as she promises; and so he may ask her just that. If Mary does as van Fraassen suggests and insists that, as a person of integrity, her word should be trusted, Ibrahim would have good reason to worry. After all, she has tried and failed before. Mary may well be a person of integrity, but something more is needed if Ibrahim is to have any reasonable confidence that Mary will actually quit smoking this time. And if Ibrahim presses her again, and she follows Marušić's advice and simply refuses to adopt the objective perspective on her behavior, then Ibrahim has very good reason not to expect that this time will be different.

Just consider what Mary could have said to Ibrahim instead of simply refusing to examine her behavior from the objective point of view. If she said that she believes that this time will be different because making the promise will give her added reason to quit, as well as added resolve, then this, at least, would be some reason to think that this time might be

different. If she explains that she has tried a similar strategy with other promises which were difficult to keep and that this often had salutary results, then Ibrahim might have even greater reason to think that this time might well be different. But both of these remarks would require Mary to take an objective perspective on her behavior rather than just the participant perspective.

Notice, as well, that if Mary does adopt the objective perspective, and she finds some real reasons for optimism, then this in no way undermines her agency. An objective view of her own behavior could, of course, undermine her agency if, for example, the objective view forces her to recognize that she is powerless to change her behavior given the nature of her addiction to tobacco and the features of her own character. But in such a case, it was these things about Mary which undermined her agency, not her taking the objective perspective.

And if Mary is, instead, more fortunate and an objective perspective reveals that she is able to withstand the temptations of tobacco this time, then the objective perspective will not undermine her agency; it will allow her to recognize the reasons for which she will be able to follow through on her promise. Admittedly, what conditions will need to be satisfied for the objective perspective to make room for agency will depend on whether compatibilism or incompatibilism is true. But either way, the objective perspective could, in principle, reveal that those conditions are met and that when Mary promises, she will indeed follow through as a result of the exercise of her own agency. So Marušić should not worry that the objective perspective would, of necessity, undermine Mary's view of herself as a free agent.

The attempt to find reasons to believe that one will follow through on one's promises and resolutions which somehow bypass the objective perspective is therefore misguided. An objective perspective doesn't serve to undermine agency. Agency can, of course, be undermined by facts about the agent's psychology in conjunction with facts about the agent's situation, but then the objective perspective merely reveals what was independently true: that the agent is in no position to do as they promise or resolve. And when the facts are more congenial and the agent is actually in a position to do what was promised or resolved, the objective perspective will make this plain to the agent, which will surely play no role in undermining agency. Those who make promises and resolutions have no reason to resist taking an objective perspective on the facts of their lives.

Conclusion

Strawsonian epistemologists are thus mistaken in thinking that there are benefits to be had by taking a participant perspective on testimony, self-knowledge, or the activities of promising or resolving and refusing to occupy the objective perspective. To the extent that the participant perspective differs from the objective perspective, either by leaving things out which the objective perspective includes or by virtue of elements which are incompatible with an objective view of one's self and one's situation, the participant perspective is thereby inferior to an objective view of matters. Strawson himself, and Strawsonian epistemologists, exaggerate the extent to which an objective perspective will result in alienation, either from others, from one's own beliefs, or from one's agency. Unsurprisingly, departures from the objective perspective are not conducive to believing as one ought.

Acknowledgments

I am grateful for comments on a previous version of this paper to Nathan Ballantyne, David Dunning, the participants at the Fordham Conference, and, especially, Berislav Marušić.

References

Anscombe, G. E. M. (1957). *Intention*. Cornell University Press.

Baron, M. (1984). The alleged moral repugnance of acting from duty. *The Journal of Philosophy*, *81*(4), 197–220.

Brownstein, M., & Saul, J. (Eds.). (2016). *Implicit bias and philosophy* (2 vols.). Oxford University Press.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford University Press.

Cassam, Q. (2015). *Self-knowledge for humans*. Oxford University Press.

Christensen, D. (2007). Epistemology of disagreement: The good news. *The Philosophical Review*, *116*(2), 187–217.

Evans, G. (1982). *The varieties of reference*. Oxford University Press.

Fricker, E. (2006). Second-hand knowledge. *Philosophy and Phenomenological Research*, *73*(3), 592–618.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*(1), 1–14.

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press.
Hinchman, E. (2014). Assurance and warrant. *Philosophers' Imprint*, *14*(17), 1–58.
Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus, and Giroux.
Kornblith, H. (1999). Distrusting reason. *Midwest Studies in Philosophy*, *23*(1), 181–196.
Kornblith, H. (2012). *On reflection*. Oxford University Press.
Kornblith, H. (2013). Naturalism vs. the first-person perspective. *Proceedings and Addresses of the American Philosophical Association*, *87*, 107–126.
Kornblith, H. (2015). The role of reasons in epistemology. *Episteme*, *12*(2), 225–239.
Kunda, Z. (1999). *Social cognition: Making sense of people*. MIT Press.
Marušić, B. (2015). *Evidence and agency: Norms of belief for promising and resolving*. Oxford University Press.
McMyler, B. (2011). *Testimony, trust, and authority*. Oxford University Press.
Michaelian, K. (2010). In defense of gullibility: The epistemology of testimony and the psychology of deception detection. *Synthese*, *176*(3), 399–427.
Moran, R. (2001). *Authority and estrangement*. Princeton University Press.
Moran, R. (2005). Getting told and being believed. *Philosophers' Imprint*, *5*(5), 1–29.
Nagel, T. (1997). *The last word*. Oxford University Press.
Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.
Paul, L. A. (2014). *Transformative experience*. Oxford University Press.
Strawson, P. F. (1997). Freedom and resentment. In D. Pereboom (Ed.), *Free will* (pp. 119–142). Hackett. (Original work published 1962).
van Fraassen, B. (1984). Belief and the will. *The Journal of Philosophy*, *81*(5), 235–256.
Willaschek, M. (2013). Strawsonian epistemology: What epistemologists can learn from "Freedom and Resentment." *Grazer Philosophische Studien*, *87*, 99–128.
Williams, B. (1981). *Moral luck: Philosophical papers 1973–1980*. Cambridge University Press.
Wilson, T. (2004). *Strangers to ourselves: Discovering the adaptive unconscious*. Harvard University Press.

12

Attitude Psychology and Virtue Epistemology

A New Framework

Alessandra Tanesini

With a few exceptions, virtue ethicists and epistemologists are committed to the empirical reality of virtues and vices.[1] They have, however, encountered difficulties in identifying the psychological constructs that underpin these traits of character.[2] Worse still, some critics, relying on work in empirical psychology, have argued that the character traits postulated by virtue theorists are fictitious.[3] In this chapter I articulate a novel account of the psychological underpinnings of character virtues and vices. In my view character virtues and vices are underpinned by clusters of attitudes serving distinctive functions.[4] Attitudes are acore construct of social psychology; they are summative evaluations of objects based on relevant evaluative beliefs, affects, and memories of past behaviors. I have previously defended the thesis that some intellectual virtues and vices are to be understood in terms of attitudes (Tanesini, 2018a).[5] Here, I expand on that account to highlight the fecundity of this approach when accounting for the vices of intellectual arrogance and servility.

This chapter has three main goals. The first is to develop a novel theory of the psychology of character virtues and vices according to which they are underpinned by attitudes. The second is to defend this account by showing

[1] Driver (2001, 2016) is the most notable of these exceptions.

[2] Possible candidates include units of the cognitive affective personality system (CAPS) (Mischel & Shoda, 1995). For accounts of virtues as CAPS traits, see especially Snow (2010) and Russell (2009).

[3] This is the so-called situationism challenge, which has been articulated by Doris (2002), Harman (2000), and, with regard to epistemic virtues, Alfano (2012).

[4] My focus here is on a small number of intellectual character vices. More would need to be said to defend fully the claim that every virtue and vice is underpinned by attitudes.

[5] Webber (2015) has also offered an account of some moral virtues in terms of attitudes.

Alessandra Tanesini, *Attitude Psychology and Virtue Epistemology* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press. © Oxford University Press 2022.
DOI: 10.1093/oso/9780197636916.003.0013

that attitudes possess the most significant features of virtues and vices. The third is to highlight the fecundity of this theory by using it to explain the characteristic manifestations of epistemic arrogance and servility. My argumentative strategy proceeds as follows. In the first section I provide an overview of some key features of moral and intellectual character virtues and vices. In the second section I explain the notion of attitude and describe its principal aspects. My focus is on attitude's object, content, structure, function, and strength. In the third section I show that attitudes possess the key features attributed to character virtues and vices. Attitudes therefore emerge as a likely candidate for the construct that underpins character traits. In the fourth section I provide further indirect evidence for my view. Using the theory that vices are clusters of attitudes, I develop detailed accounts of the nature and manifestations of two intellectual character vices: epistemic arrogance and epistemic servility or obsequiousness. Its ability to throw light onto the complexities of these character traits is a further indication of the power of the theory.

Intellectual Virtues and Vices

My aim in this section is to provide a broad description of some key features of intellectual character virtues and vices. What follows is a list of nine properties of intellectual virtues and vices. The first five qualities are common to all character traits, including some, such as being tidy, that are not either a virtue or a vice. The final four properties instead are distinctive of virtues and of vices.

1. Virtues and vices are components of a person's character. Qualities like courage or perseverance are deep features of individuals. They contribute to defining who people are; they express people's deeply held commitments and values (Annas, 2011, p. 9; Miller, 2013).
2. Virtues and vices are stable over time and consistent across situations. That is, they are psychological qualities that are predictive of behavior across a broad range of situations and over extended periods of time. Individuals who display the behavior characteristic of a virtue or vice inconsistently or only occasionally are not thought to possess the relevant character trait (Miller, 2014).

3A. Virtues and vices are often described as *multi-track dispositions* to act, feel, and judge in characteristic ways. These dispositions are said to be multi-track because they are triggered by diverse stimuli, responding to each in distinctive ways. For example, individuals who are courageous are disposed to act bravely in heterogenous situations such as saving a person from a fire or standing up to a bully. Further, what the courageous person is disposed to do varies in accordance with the situation they face (see Webber, 2006, 2015). While numerous virtues and vices have this structure, others do not.

3B. There are other virtues and vices for which the multi-track dispositional model is not an equally good fit. For instance, some virtues motivate their possessors to bring about the circumstances which demand their exercise. Thus, the generous person may seek to bring about opportunities to express their generosity. In addition, some virtues, such as humility, integrity, and perseverance, fit practically any circumstances and thus require near continuous exercise (see Webber, 2015).

4. Virtues and perhaps also vices, are intelligent because they are flexible and responsive to reason (Annas, 2011; Snow, 2010). These points are often couched in the vocabulary of skill. For example, the person who is generous must be able to discriminate whether a friend in difficult economic circumstances would welcome a monetary gift or whether they would find such a gift humiliating. The generous individual can "read" the situation and understand what is, in those circumstances, the right thing to do. These abilities to be sensitive to subtle cues and to understand the demands posed by a situation are complex skills that need to be learned and refined. Arguably, some vices also require the possession of skills. For instance, to be effective a malevolent individual must be able to appraise which situations offer the best opportunity to harm other people.

5. Often virtues and vices are said to have emotional components, or at least characteristic emotions associated with them (Zagzebski, 1996). For instance, individuals who are vain typically experience envy when confronted with other people's achievements. Those who are generous usually experience feelings of benevolence toward others. Arrogant individuals are often angry, while those who are timid are frequently fearful.

So far, I have described virtues and vices as psychological qualities of individuals that are constituents of their character, are intelligent and may involve some skills, are stable over time, and are manifested consistently across situations. They are also often associated with emotions or may include an emotional component. I have added that they are often identified with dispositions because some of them are only manifested when triggering circumstances obtain. Plausibly, however, there are several other psychological qualities that possess some or all of these features but that are neither virtues nor vices. Among these one may include dispositions to be neat and tidy or to be lively, enthusiastic, and extroverted (Annas, 2011, p. 101). It is generally agreed that these character traits or temperaments are not virtues or vices because they are not related to goodness or badness in the relevant ways.

There are numerous, and often conflicting, philosophical accounts of the nature of the relation of virtue to goodness. I shall not take a stand on this issue here. Instead, I consider four features that are usually taken to pertain to virtues and can be used to explain their goodness.

6. Virtues have motivational components. According to this view good motivations are elements of virtuous character traits (for a review, see Battaly, 2015, pp. 15–18). Zagzebski (1996) argues that all intellectual virtues share a motive of love of truth or cognitive contact with reality, while each is individuated by its distinctive proximate motivation.
7. Virtues are constituents of human flourishing or well-being. So understood, virtues are those character traits whose presence ensures that their possessor flourishes or leads a good human life. Intellectual virtues in particular would be those psychological qualities whose presence secures the flourishing of epistemic agents qua epistemic agents.
8. Virtues in ordinary circumstances lead reliably to good effects or good outcomes. It is not implausible to think that virtue must be effective. Hence, for example, generosity should result in an improved situation for the targets of generous acts. Instead, the person who tries to be helpful but repeatedly misjudges the situation and thus fails to provide any assistance is well intentioned without being virtuous. So conceived, virtue requires some level of success in one's endeavors. Intellectual

virtues, in particular, are plausibly thought of as traits which are truth- or knowledge-conducive (Zagzebski, 1996, pp. 176–194).[6]

9. Virtues are admirable qualities of agents. Virtues are good psychological features of human beings. But virtue is not merely something that is desirable to have; it is also some kind of achievement whose possession makes the virtuous person admirable (Zagzebski, 2015).

The idea that virtues reliably lead to good effects and/or have good motivations provides a way to distinguish between moral and intellectual virtues. Moral virtues include intrinsically good moral motivations or lead to morally good effects. Epistemic virtues, instead, comprise a motive of love, care, or concern for epistemic goods such as truth, knowledge, or understanding. They may also be reliably conducive to the acquisition and preservation of these goods (Driver, 2003).

While a plausible case can be made that virtues are character traits that include good motivations, contribute to human flourishing, reliably produce good effects, and are admirable, it is at best unclear whether vices must possess the opposite characteristics. I have argued elsewhere that intellectual vices comprise non-instrumental aversion to things which are epistemically good in themselves (Tanesini, 2018b). But this is a minority view; Cassam (2015) and Crerar (2017) have claimed, for example, that intellectually vicious individuals may be driven by epistemically good motivations. In what follows, I sidestep this issue and adopt Cassam's definition of intellectual character vices as psychological qualities that are constituents of the character of individuals and are obstacles to effective—that is, knowledge-conducive—and responsible—that is, careful and guided by the evidence—inquiry (Cassam, 2016). These vices plausibly include closed-mindedness, dogmatism, intellectual arrogance, vanity, and obsequiousness.

Attitudes

My concern in this section is with a core construct of social psychology: attitudes. These are summary evaluations of objects. They may be thought of as likes or dislikes for particular things. For instance, I have a

[6] I set aside here the position in epistemology known as virtue reliabilism that identifies intellectual virtues with cognitive faculties such as perception or memory. This position was first articulated fully by Sosa (2007).

strong aversion to liquorice. I dislike the stuff and try to avoid it if it crosses my path. My dislike of liquorice is a negative attitude toward it. It has a distinctive affective component of physical revulsion, and it guides my aversive behavior. In what follows I summarize some features of attitudes, to which I return in the subsequent section to substantiate my claim that some character virtues and vices are underpinned by clusters of attitudes directed at one's cognitive process and abilities, contributing to one's self-concept.

The main aspects of attitudes are their object, content, structure, and function. In what follows I explain these features before turning to a description of some properties of attitudes such as strength, certainty, centrality, and extremity that moderate the effects of attitudes on information processing and on attention. I also briefly address some of the issues surrounding measures of attitudes.

A. Attitude *object*. The object of an attitude is what the attitude evaluates. It can be a concrete particular such as a person's fountain pen, a kind of thing (e.g., dogs), an abstract entity or a value (e.g., justice or equality), or even a social group (e.g., university students) (Banaji & Heiphetz, 2010; Maio & Haddock, 2015).
B. Attitude *content*. This is the informational basis from which the attitude is derived. The attitude itself is the result of weighing up the positive and negative considerations relevant to the object that are conveyed by the attitude content. Thus, the attitude is a summary of its informational content and functions as a cognitive shortcut. Instead of needing to reconsider afresh each time whether one likes or dislikes a given object, a subject can access the relevant attitude to guide their behavior. Hence, attitudes save us time and cognitive effort (Banaji & Heiphetz, 2010; Fazio & Olson, 2007).

The informational content of attitudes is usually thought to include components of three different kinds: The cognitive elements include all of the evaluative beliefs one holds about the target object, the affective elements comprise all of the emotions and feelings one has about it, and the behavioral components are past behaviors and experiences regarding the object which are remembered by the subject (Maio & Haddock, 2015, pp. 29–32). Taking again my dislike of liquorice as an example, the attitude content includes, among other things, my beliefs about the taste of liquorice and its staining

qualities, my feelings of disgust, and my memories of having been sick as a child while in the presence of silver-colored liquorice lozenges.

The contents of attitudes change as we acquire new information about their objects or have novel experiences of them. Such changes in attitude contents bring about changes in the attitudes themselves. Hence, attitudes are responsive to rational considerations since they update in the light of novel information and new experiences. However, the reason responsiveness of attitudes is somewhat patchy, as I explain below (see "Intellectual Vices, Attitude Function, and Biases").

C. Attitude *structure*. The structural features of the informational basis or content of an attitude contribute to determining the attitude's influence on behavior. Recently, the view that the information included in the attitude content is structured along two dimensions has become dominant. Thus, we should not think of the attitude's valence as an aggregate evaluation somewhere on a spectrum from very favorable to neutral to very negative. Instead, positive and negative elements are aggregated separately, generating attitudes whose valence can be represented as a point in a Cartesian coordinate system where one axis measures levels of positivity and the other levels of negativity (Maio & Haddock, 2015, pp. 39–41). For example, a subject may possess an attitude toward chocolate which is both very positive and very negative. This ambivalent attitude may result from an informational basis that includes negative beliefs based on chocolate's caloric content and positive ones concerning its health benefits but also from positive and negative feelings toward chocolate as well as pleasant and unpleasant memories. Usually attitudes that are ambivalent are more subject to situational influences than non-ambivalent attitudes. Hence, one may respond positively or negatively to the object depending on which of its features are particularly salient on a given occasion (Bell & Esses, 2002).[7]
D. Attitude *function*. Attitudes are formed, modified, and maintained in order to satisfy human needs (Maio & Olson, 2000a). These needs

[7] This kind of ambivalence is known as *potential ambivalence*. It is to be distinguished from *felt ambivalence*, which refers to a subject's feelings of tension about the attitude object (Maio & Haddock, 2015, p. 41). I shall return to felt ambivalence when I discuss intellectual servility (see "Intellectual Vices, Attitude Function, and Biases").

individuate the function or functions served by the attitude. Due to difficulties in measurement and taxonomy, the functional approach to the study of attitudes is not universally shared. I adopt it here because the classification of attitudes in terms of their functions is especially fruitful when explaining the effects of attitudes on information processing and other inquiry-relevant cognitive activities.

There are several taxonomies of attitude functions currently in use. Although these are different, there are significant areas of agreement and overlap among them. Broadly speaking, six functions have gained widespread acceptance: object appraisal, knowledge, instrumental, ego-defensive, social adjustive, and value expressive. The object appraisal function is defined somewhat trivially as the function, shared by all attitudes, that serves the need to evaluate (Fazio, 2000). I mention it here largely to set it aside in favor of another function which is thought to be one of its constituents: knowledge (Katz, 1960). Attitudes are classified as serving this function if they are formed and revised to satisfy the need to make sense of the world. This point could be couched in motivational terms (Marsh & Julka, 2000). When the motive of having an accurate account of the target guides the formation and revision of an attitude, that attitude is said to have a knowledge function.

Instrumental attitudes satisfy the need to maximize rewards. *Ego-defensive* attitudes are those that are formed in response to a need to defend the ego against real or presumed threats. Those attitudes that respond to the need to belong to one's elective social group are said to serve a *social-adjustive* function. Finally, attitudes that satisfy the need to give expression to one's values have a *value-expressive* function (Maio & Haddock, 2004; Maio & Olson, 2000b). The target object of a value-expressive function need not itself be a value. For example, I have a positive attitude toward my walking boots. They remind me of the great outdoors and of feelings of freedom, autonomy, and awe. Thus, it is plausible to conclude that in my case this attitude satisfies the need to give expression to a set of values.

A single attitude may serve more than one function. For instance, it is plausible that ambivalent attitudes toward chocolate of the kind I have described satisfy more than one need. A person's attitude toward chocolate may be negative because avoiding the stuff contributes to satisfying the need for social acceptance (since indulgent eating is frowned upon) but also positive since eating chocolate satisfies a hedonistic need (Maio & Olson, 2000a, pp. 429–430).

As should be clear from what I have said so far, the study of attitudes is concerned with investigating psychological differences between agents. That is because the contents, function, valence (positive or negative), and structure of attitudes vary from person to person. The role played by attitudes in guiding inquiry-relevant behavior is moderated by some key features of the attitudes themselves. I focus here on four features which, confusingly, have all been described as attitude strength.

The first of these is attitude strength understood as attitude accessibility (Fazio, 2000). The kind of strength that is at issue here is that of the associative link between the two components of the attitude itself: the representation of the target object and a valence. The stronger the association, the more likely it is that anything that triggers the representation of the object will also activate the valence (Fazio et al., 1986). So understood, attitude strength is a measure of accessibility. In what follows, unless stated otherwise, I shall use the term *attitude strength* in this interpretation.

Often, however, the term is deployed to signify some rather distinct aspects of attitudes. One is attitude extremity. Attitudes are said to be extreme when someone is very positive or very negative about the target object. My negative attitude to liquorice, for example, is extreme. While there may be correlations between attitude strength and extremity, the two are conceptually distinct and have different effects on information processing (Brannon et al., 2007). *Strength* is also used to refer to attitudes that are important or central to the subject because they contribute to defining the self and are close to one's values and commitments (Clarkson et al., 2009; Zunick et al., 2017). Quite often, *strength* refers to attitude certainty, which concerns the subject's confidence in their attitudes. This notion of attitude certainty, as should be clear, is itself ambiguous between two different notions. The first is *correctness*, which refers to the subject's certainty that their attitude is accurate or correct. The second is *clarity*, which measures the subject's certainty that a statement expresses their attitude (Petrocelli et al., 2007). Attitude certainty as correctness is opposed to feelings of doubt about the rightness or truth of one's attitude. Often, all or some of these diverse kinds of strength are measured before being aggregated. Hence, empirical work on the moderating role of attitude strength on information processing often treats strength as an aggregate measure of several distinct factors.

I conclude this section by mentioning another aspect of attitudes that relates to their measurement. While there are many different measures of attitudes, these broadly divide into two distinct kinds: explicit and implicit.

Attitudes are measured explicitly or directly by means of questionnaires and self-reports, often recording agreement or disagreement with a statement as ranked in a Likert scale (Maio & Haddock, 2015, pp. 10–14). They are measured implicitly or indirectly by measuring the speed and accuracy of responses in implicit association tests or by means of evaluative priming or still by other means that only measure the attitude indirectly and without the subject's awareness of what is being measured (Maio & Haddock, 2015, pp. 14–21). Explicit and implicit measures of attitudes dissociate. This fact alone, however, does not establish that they tap in to different constructs. It is possible that implicit measurements' divergence from explicit ones is the result of depriving people of the opportunity and motivation to examine their attitudes before acting on them (Fazio & Olson, 2003, pp. 303–305).

Vices and Virtues as Cluster Attitudes

In the first section I enumerated several features that characterize virtues and vices. I claimed that these are generally thought to be components of people's characters: They are stable over time and across situations, are responsive to rational considerations, have an emotional component, and can be thought of as dispositions. I added that virtues are characterized by their connection to the good because they generally comprise good motivations, contribute to human flourishing, and reliably produce good outcomes. In addition, virtues are generally thought to be admirable. Vices, on the other hand, are defects that may involve an aversion to the good. Intellectual vices are obstacles to inquiry that is responsibly carried out and effective in the production and transmission of knowledge.

Once these feature of virtues and vices are highlighted, the overlap with the properties of attitudes can be made salient. Some attitudes are part of people's characters. Such attitudes are said to be central or important to the person. The same point can be made in functional terms. Some attitudes satisfy the need to express one's values; when these values are central to the self-concept, the attitude is part of the person's character. For instance, for some people egalitarian values are deeply important. These values may have led them to form attitudes about career choice, donating to some charities, and political affiliation. These people may work in non-profit organizations, be left-leaning politically, and donate to charity a percentage of their salary every month. It seems plausible to think of the attitudes which are at the root

of these behaviors as being part of an individual's self-concept and thus of their character.

When attitudes are strong, and when they are central or held with certainty, they are stable over time and consistent across situations (Luttrell et al., 2016; Petty et al., 1997, p. 634). Functions and structure interact with attitude stability. Attitudes that are ambivalent, for example, are less stable since they are subject to situational factors (Crano & Prislin, 2006, p. 365; Luttrell et al., 2016). Depending on the context, different elements of the attitude content may be activated, thus leading to varied responses in diverse situations. Attitude function is also relevant since, for example, attitudes serving the need for social adjustment are more sensitive to situational variation than value-expressive or ego-defensive attitudes (Levin et al., 2000). This is not surprising since individuals with a pressing need to be accepted in a group may adjust their opinions to what they think is the majority view in their elective social group.

Individuals can change their attitudes in response to novel experiences or when acquiring further relevant information. Attitudes can update rationally but are at times resistant to new information. As I discuss below (see "Intellectual Vices, Attitude Function, and Biases"), the literature on attitude change reveals that individuals do not always revise their attitudes in the light of the evidence since their responses are often motivated by goals other than accuracy (Levin et al., 2000). That said, the argumentative quality of persuasive messages aimed at changing people's view is usually a significant factor in determining whether subjects change their attitudes. Further, there are reasons to believe that even attitudes revealed by implicit measure are involved in inferential processing in some way (see Levy, 2015). In short, attitudes are intelligent. In addition, strong attitudes can direct visual attention to attitude-relevant aspects of the situation (Fazio, 2000). This feature of strong attitudes suggests that they play a role in the skillful navigation of the environment.

Attitudes have an emotional element since their contents include affective and emotional components. These constituents are extremely significant because affective factors are often the best predictors of subjects' attitudes (Maio & Haddock, 2015, p. 37).

Finally, attitudes can be, and have been, largely conceived as enduring dispositions that influence a broad range of behaviors (Ajzen, 1987, p. 1). In psychology at least, even though research on attitudes and research on personality traits have generally proceeded independently of each other, it is

generally presumed that these two kinds of construct are similar. There are two main differences: Attitudes are directed at objects and are always evaluative (Sherman & Fazio, 1983). For instance, being aggressive or being extroverted would be personality traits, while disliking people or liking social gatherings would count as attitudes.

Virtues and vices are in this regard closer to attitudes than to personality traits. Virtues and vices are always evaluative since they involve appraisals of the current situation as demanding a specific kind of response. Thus, virtues are akin to attitudes in so far as they issue evaluations of an object or a situation in the light of the person's motives. For example, a person who is open-minded may respond to a challenge to their views by listening carefully, considering the possibility that they are mistaken. Hence, to be open-minded involves evaluating other people's views in the ways which are characteristic of open-mindedness. Humility too involves the development of attitudes toward one's own limitations and achievements that evaluate them in a manner that is consonant with their objective qualities (Tanesini, 2018a). Similarly, intellectual arrogance consists in the formation of clusters of attitudes toward one's own achievements and shortcomings and those of other agents that are predictive of a range of aggressive, closed-minded, and dismissive behaviors (Tanesini, 2016b).

To summarize, attitudes possess key features of character traits and thus are a plausible candidate for being the type of mental state that underpins virtues and vices. Clearly, not all attitudes can fulfill this role. To do so, attitudes must be central or important to the subject and strong in the sense of being highly accessible. In addition, attitudes must bear a close relation to the good or the bad.[8] We can think of moral virtues as clusters of strong attitudes toward a range of objects and situations that contribute to the flourishing of those who have them, reliably lead them to bring about morally good outcomes, and include good motives and emotions in their informational bases. These points can be couched in the vocabulary of attitude function. Virtues are underpinned by strong attitudes serving value-expressive functions and are such that the values they express actually promote human flourishing.

[8] To my knowledge, this feature of some attitudes has not so far figured highly on the social psychological research agenda.

Intellectual virtues and vices, with which I am concerned here, are also underpinned by clusters of attitudes.[9] Crucial for intellectual virtues are attitudes that include a motivation to pursue epistemic goods and that lead reliably to their acquisition. These are goods, such as truth and knowledge, that contribute to the intellectual flourishing of those who acquire them. Prominent among these attitudes are those that are positive and directed toward the epistemic goods themselves. These points can be expressed using a functional approach to the categorization of attitudes. Intellectual virtues are underpinned by strong and central attitudes that have been formed and sustained by the need for knowledge, which is to say the need to make sense of the world. These include attitudes to epistemic goods serving knowledge and perhaps value-expressive functions. Contrary to virtues, intellectual vices are obstacles to the intellectual flourishing of those who possess them. They may also include aversive motivations to epistemic goods and in some cases, such as epistemic malevolence (Baehr, 2010), negative attitudes toward epistemic goods or other people's share of them.

Intellectual Vices, Attitude Function, and Biases

In this section I rely on the account offered in the previous section to illustrate some of the ways in which intellectual vices are obstacles to effective and responsible inquiry. My focus here is on two vices opposed to intellectual humility: intellectual arrogance and intellectual servility or obsequiousness. I argue that intellectual arrogance is underpinned by clusters of strong ego-defensive attitudes that are typical of what is known as high defensive self-esteem. Servility is, instead, an expression of damaged self-esteem, consisting of a cluster of attitudes serving a social-adjustive function. After a description of the psychological architecture of these vices, I highlight their biasing influence on reasoning and on information seeking and processing.

Intellectually arrogant behaviors are varied. They include being condescending and dismissive of other people's opinions and being full of oneself and feeling intellectually superior. Arrogant individuals often interrupt others or do not allow them to put a word in edgeways. Often, they react angrily to what may seem sensible criticisms of their views. In general, those

[9] I say that virtues and vices are underpinned by attitudes because attitudes are the psychological states that ground the dispositions constitutive of virtuous and vicious character traits.

who are arrogant claim for themselves special entitlements that they deny to other people. So conceived, intellectually arrogant individuals have a high opinion of their intellectual abilities and, in general, of their competence. They also display a tendency to compare themselves favorably to others. Further, this feeling of superiority and sense of one's own excellence is quite central to their conception of themselves.

These features of intellectual arrogance all indicate that arrogant behaviors may flow from strong positive attitudes about the self and one's own intellectual abilities. These attitudes are likely to be strong in people who frequently express these behaviors. Given the association between attitude strength and centrality to attitude certainty (Tormala & Rucker, 2007, p. 471), these individuals are also likely to be certain about their views both in the sense of feeling that they know their own minds and in the sense of thinking that their views are correct. However, it is unlikely that high self-esteem and positive attitudes toward one's own abilities are by themselves the cause of arrogance. It seems possible to have a good opinion of oneself and of one's intellectual competence while being quite respectful of other people and even humble, or at least justly proud, of one's achievements.

In my view the difference between those who are arrogant and those who are properly proud of their intellectual track record lies in the function served by their strong and positive attitudes. In the person who is arrogant, high self-esteem and positive attitudes about one's own intellectual abilities serve an ego-defensive function. This individual has formed positive attitudes about their own abilities in response to a need to feel good about themselves. Thus, they evaluate positively some of their qualities because these satisfy a need for self-enhancement. Consider, for instance, a person who is particularly arrogant about their problem-solving ability. This person would have a strong positive attitude about their competence in problem solving. They would also think of this attitude as important and as part of their self-conception. The attitude serves an ego-defensive function if the person has developed, and maintains, the attitude because of its effectiveness in preserving high self-esteem. It is possible that this individual is objectively a skillful problem solver; hence, the arrogant individual could in principle be accurate in their self-assessment, although this is unlikely. Be that as it may, the significant point is that the attitude is not responsive to the person's actual ability; instead, what the attitude tracks is effectiveness in ego defense. Self-enhancement is the motive for forming the attitude and for continuing to hold it (Bosson et al., 2003).

There is a substantial body of empirical research on individuals who are particularly prone to possessing defensive attitudes These people have high self-esteem, when this attitude is measured explicitly. But they suffer from low implicitly measured self-esteem (Jordan et al., 2003). They are, thus, said to have discrepant self-esteem. It is this underlying lack of self-esteem, which is recorded by implicit measures, that may be at the root of the defensive need of self-enhancement. This defensiveness results in positive explicitly measured self-esteem satisfying the need to repel threats to the ego. There are robust results indicating that people with defensive high self-esteem behave in ways which are characteristic of arrogance. These include a propensity to anger (Schröder-Abé et al., 2007) and aggression (Kirkpatrick et al., 2002), a tendency to boast (Olson et al., 2007) and to respond arrogantly to threats (McGregor et al., 2005), a disposition to self-enhance (Bosson et al., 2003), and heightened defensiveness (Haddock & Gebauer, 2011). These findings provide significant evidence in favor of the hypothesis that intellectual arrogance is underpinned by those defensive positive attitudes about one's abilities and competencies that are characteristic of individuals possessing a defensive high self-esteem.[10]

At the opposite end of the scale from arrogance lies intellectual obsequiousness or servility. Intellectual obsequiousness is manifested in a tendency to self-deprecate, to attribute one's successes to luck or the ease of the task and one's failures to incompetence (depressive attributional style). Intellectually servile people are beset by feelings of inferiority, lack of confidence, anxiety, and shame because of their worthlessness. What makes these individuals obsequious is their tendency to respond to their sense of inadequacy by deferring to other's people views, by trying to seek the approval of powerful members of their social group, and by attempting to ingratiate themselves to individuals who may bully them or treat them disrespectfully.

These features of obsequiousness suggest that it may result from negative attitudes toward the self and one's own cognitive abilities and level of competence that serve a social-adjustive function. That is, intellectually obsequious individuals are driven by the need to be accepted within a social group to which they wish to belong. Perhaps due to ill fortune or to prejudice, these individuals find it hard to gain social acceptance as equal members of the group. Instead, they adopt the strategy of making others feel superior to them

[10] I have offered a more detailed description of arrogance in Tanesini (2016a) and a defense of this hypothesis in Tanesini (2016b).

to gain their acceptance (Vohs & Heatherton, 2004). These individuals have been described as possessing a damaged self-esteem because they have low self-esteem when this is measured explicitly but measure high in implicitly measured self-esteem (Vohs & Heatherton, 2004). Although less is known about these subjects than about their arrogant counterparts, there is good evidence that they possess those tendencies to ingratiate, to adopt a depressive attributional style, and to be riven by anxiety and self-doubt, which are the characteristic manifestations of intellectual servility (Tanesini, 2018c).

Both intellectual arrogance and servility are intellectual vices. Intuitively, those who are arrogant dismiss other people's views and thus fail to learn from them. In addition, their self-confidence may cause them to become closed-minded and dogmatic and therefore terminate their inquiries too soon without having considered sufficient evidence. People who are obsequious, on the other hand, may be too ready to defer to the views advanced by powerful individuals without considering whether these opinions are supported by the evidence. These same individuals may also have a tendency never to form firm beliefs due to their propensity to self-doubt and anxiety. In what follows, I show how research on attitudes throws some light on these phenomena.

Defensive high self-esteem and, in general, a tendency to form strong ego-defensive attitudes has been shown to bias what one pays attention to. This phenomenon, which is known as the *selective exposure effect*, is the propensity to notice and seek only information that confirms one's pre-existing attitudes (Maio & Haddock, 2015, pp. 56–60). Such a propensity is tantamount to a congeniality effect or confirmation bias. While there is some evidence that all attitudes bias our information-seeking behaviors, not all of these biases are epistemically bad in the sense of making one's beliefs less accurate. For example, attitude importance and accessibility may predict increased attention to attitude-relevant information (Maio & Haddock, 2015, p. 58). In this sense accessibility biases attention because it increases the likelihood that it is drawn to one object rather than another. This bias, however, may well be rational because it may promote the accuracy of a person's object-relevant beliefs. Here, I am not interested in biases of this kind but only in those that are obstacles to knowledge-conducive inquiry.

There is empirical evidence that a motivation to defend one's pre-existing attitudes unsurprisingly amplifies confirmation biases (Hart et al., 2009). It is plausible that this defensive motivation is of a piece with how much of the ego is invested in the attitude. Thus, ego-defensive attitudes would be

generally associated with a defensive motivation (Levin et al., 2000). People whose attitudes are defensive would be especially prone to focusing their attention only on information that confirms their views, while paying less attention to disconfirming evidence. There is substantial empirical evidence confirming this hypothesis. People whose high self-esteem is defensive exhibit a heightened tendency to defend their own attitudes (McGregor & Marigold, 2003). Hence, we would expect these individuals to be more prone than other people to congeniality effects.

There is additional evidence that certainty in the correctness of one's attitudes also amplifies confirmation biases (Knobloch-Westerwick & Meng, 2009). Since high-defensive self-esteem is also positively correlated to attitude certainty (McGregor & Marigold, 2003), we should expect these individuals to be particularly susceptible to biases because of their unwillingness to consider counter-attitudinal information and because their defensiveness leads them to dismiss others' views when they conflict with theirs. These expectations have received empirical confirmation. People whose high self-esteem is defensive have a propensity to react badly to negative feedback by derogating the views of out-group members (Jordan et al., 2005). They exhibit higher levels of prejudice toward members of other ethnic groups (Jordan et al., 2005), they are prone to higher levels of self-deception in general than those whose high self-esteem is congruent (Jordan et al., 2003), and they have a propensity to overestimate the extent to which other people agree with their views (McGregor et al., 2005).

Defensive attitudes thus have biasing effects on searching for and processing of information. For this reason, they are especially resistant to rational update in the light of novel information (Levin et al., 2000). While there is no consensus on the exact details, there is broad agreement that individuals whose motives are defensive will be particularly alert to information that speaks to their concerns.[11] Although they may process such information more systematically, their processing is likely to be biased against changing their views. It is this resistance to attitude change that makes it plausible to think that individuals whose high self-esteem is defensive are closed-minded and dogmatic (Hart et al., 2009, p. 558). In short, the identification

[11] There are several competing models of the effects of attitude function on persuasion. These predict function-matching effects. Individuals are more interested in information that speaks to the needs served by their attitudes. Therefore, they are disposed to process more deeply, but not necessarily in a less biased way, those messages whose content matches the function of their attitudes. See Petty and Cacioppo (1986) for the elaboration likelihood model and Chen et al. (1999) for the heuristic and systematic model.

of intellectual arrogance as being underpinned by high defensive self-esteem vindicates the prediction that arrogance is an intellectual vice because it promotes self-deception, confirmation biases, closed-mindedness, and dogmatism.

The effects of social-adjustive attitudes on attention and information processing are less well understood. But there is suggestive evidence linking these attitudes to the kind of deferential behavior that gets in the way of knowledge acquisition. Studies have shown that individuals who are motivated to impress others tend to accept what their conversational partners say to get along (Chen et al., 1996).[12] These individuals usually do not have stable attitudes. Instead, they change their views depending on whom they listen to. None of these opinions are stable since, having first agreed with someone when talking to them, they then often revert to what they thought earlier (Levin et al., 2000, p. 178). Hence, these individuals tend to agree with what others say rather than to examine the evidence.[13] Their inquiries, as a result, are less likely to produce knowledge.

In my view intellectual obsequiousness would be especially characteristic of those people who exhibit a damaged self-esteem. There is limited empirical research on this group of individuals. Current work could be taken to indicate that they are less defensive than others whose low self-esteem is congruent because their high implicit self-esteem buffers them against ego threats (Spencer et al., 2005). In addition, there is evidence that people whose self-esteem is damaged are especially responsive to feedback about successes and failures (Jordan et al., 2013). This responsiveness may be responsible for their perfectionist tendencies (Zeigler-Hill & Terry, 2007). It may also be taken as an indication that their self-esteem is more dependent on external circumstances and other people's opinions of their abilities. If this is right, these individuals would be disposed to having attitudes serving a social-adjustive function.

There is evidence that social-adjustive attitudes tend to be weaker and more subject to contextual influences than attitudes serving other functions (Levin et al., 2000). This is what one would expect if social-adjustive attitudes are formed with the goal of getting along with other people who may disagree

[12] This motive to be socially accepted is akin to the social-adjustive function played by some attitudes (Levin et al., 2000). The same motive has been described by some as an indirect form of self-enhancement (Brown et al., 1988).

[13] They may also exhibit high self-monitoring, understood as a disposition to match one's conduct to the social demands of the situation (see Gangestad & Snyder, 2000).

with each other. Although there is to my knowledge no research explicitly linking discrepant low self-esteem with felt attitude ambivalence, one would expect these individuals to experience doubts and internal conflict about their opinions. Since ambivalence in attitudes is associated with fearfulness of errors and lack of confidence in the correctness of one's views (Thompson & Zanna, 1995), it seems plausible that low discrepant self-esteem individuals may be especially prone to having attitudes that are ambivalent.[14] If this is right, these individuals, in addition to agreeing with others, and especially powerful others, in order to get along, may be prone to changing their minds frequently depending on what is salient in a given context (Bell & Esses, 2002). Finally, ambivalence is associated with indecision in decision-making and a propensity for procrastination (van Harreveld et al., 2009). In sum, research on attitudes throws light on the link between low discrepant self-esteem and a desire to please to be accepted. Those who possess these characteristics exhibit behaviors that are obstacles to inquiry which is responsible and effective.

Conclusion

In this chapter I have advanced a new account of virtues and vices that sees them as underpinned by attitudes. I have first highlighted some key features of these character traits. Subsequently, I have introduced the attitude construct and explained why attitudes possess the qualities traditionally attributed to virtues and vices. I have concluded that this congruence offers some support for the hypothesis that attitudes are the psychological states responsible for virtues and vices. However, the strongest argument in favor of this account is provided by its ability to throw light on the nature and characteristic expressions of specific virtues or vices. The analyses of intellectual arrogance and obsequiousness presented in the final section of this chapter exemplify the explanatory power of the theory defended here.

[14] Work suggesting an association between lack of self-clarity and low self-esteem might offer further support for these claims (see Campbell, 1990). These individuals might be unclear about which attributes best define them. This research, however, makes no distinction between congruent and discrepant low self-esteem.

References

Ajzen, I. (1987). Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology. *Advances in Experimental Social Psychology*, *20*, 1–63.

Alfano, M. (2012). Expanding the situationist challenge to responsibilist virtue epistemology. *The Philosophical Quarterly*, *62*(247), 223–249.

Annas, J. (2011). *Intelligent virtue*. Oxford University Press.

Baehr, J. (2010). Epistemic malevolence. *Metaphilosophy*, *41*(1–2), 189–213.

Banaji, M. R., & Heiphetz, L. (2010). Attitudes. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., Vol. 1, pp. 353–393). John Wiley & Sons.

Battaly, H. (2015). A pluralist theory of virtue. In M. Alfano (Ed.), *Current controversies in virtue theory* (1st ed., pp. 6–22). Routledge.

Bell, D. W., & Esses, V. M. (2002). Ambivalence and response amplification: A motivational perspective. *Personality and Social Psychology Bulletin*, *28*(8), 1143–1152.

Bosson, J. K., Brown, R. P., Zeigler-Hill, V., & Swann, W. B. (2003). Self-enhancement tendencies among people with high explicit self-esteem: The moderating role of implicit self-esteem. *Self and Identity*, *2*(3), 169–187.

Brannon, L. A., Tagler, M. J., & Eagly, A. H. (2007). The moderating role of attitude strength in selective exposure to information. *Journal of Experimental Social Psychology*, *43*(4), 611–617.

Brown, J., Collins, R. L., & Schmidt, G. W. (1988). Self-esteem and direct versus indirect forms of self-enhancement. *Journal of Personality and Social Psychology*, *55*(3), 445–453.

Campbell, J. D. (1990). Self-esteem and clarity of the self-concept. *Journal of Personality and Social Psychology*, *59*(3), 538–549.

Cassam, Q. (2015). Stealthy vices. *Social Epistemology Review and Reply Collective*, *4*(10), 19–25. http://wp.me/p1Bfg0-2na

Cassam, Q. (2016). Vice epistemology. *The Monist*, *99*(2), 159–180. https://doi.org/10.1093/monist/onv034

Chen, S., Duckworth, K., & Chaiken, S. (1999). Motivated heuristic and systematic processing. *Psychological Inquiry*, *10*(1), 44–49.

Chen, S., Shechter, D., & Chaiken, S. (1996). Getting at the truth or getting along: Accuracy- versus impression-motivated heuristic and systematic processing. *Journal of Personality and Social Psychology*, *71*(2), 262–275.

Clarkson, J. J., Tormala, Z. L., DeSensi, V. L., & Christian Wheeler, S. (2009). Does attitude certainty beget self-certainty? *Journal of Experimental Social Psychology*, *45*(2), 436–439.

Crano, W. D., & Prislin, R. (2006). Attitudes and persuasion. *Annual Review of Psychology*, *57*, 345–374.

Crerar, C. (2017). Motivational approaches to intellectual vice. *Australasian Journal of Philosophy*, *96*(4), 753–766. https://doi.org/10.1080/00048402.2017.1394334

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge University Press.

Driver, J. (2001). *Uneasy virtue*. Cambridge University Press.

Driver, J. (2003). The conflation of moral and epistemic virtue. *Metaphilosophy*, *34*(3), 367–383.

Driver, J. (2016). Minimal virtue. *The Monist*, *99*(2), 97–111. https://doi.org/10.1093/monist/onv032

Fazio, R. H. (2000). Accessible attitudes as tools for object appraisal: Their costs and benefits. In G. R. Maio & J. M. Olson (Eds.), *Why we evaluate: Functions of attitudes* (pp. 1–36). Lawrence Erlbaum Associates.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*(1), 297–327.

Fazio, R. H., & Olson, M. A. (2007). Attitudes: Foundations, functions and consequences. In M. A. Hogg & J. Cooper (Eds.), *The SAGE handbook of social psychology* (concise student ed., pp. 139–160). SAGE.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), 229–238.

Gangestad, S. W., & Snyder, M. (2000). Self-monitoring: Appraisal and reappraisal. *Psychological Bulletin*, *126*(4), 530–555.

Haddock, G., & Gebauer, J. E. (2011). Defensive self-esteem impacts attention, attitude strength, and self-affirmation processes. *Journal of Experimental Social Psychology*, *47*(6), 1276–1284.

Harman, G. (2000). The nonexistence of character traits. *Proceedings of the Aristotelian Society*, *100*, 223–226.

Hart, W., Albarracin, 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.

Jordan, C. H., Logel, C., Spencer, S. J., Zanna, M. P., Wood, J. V., & Holmes, J. G. (2013). Responsive low self-esteem: Low explicit self-esteem, implicit self-esteem, and reactions to performance outcomes. *Journal of Social and Clinical Psychology*, *32*(7), 703–732. https://doi.org/10.1521/jscp.2013.32.7.703

Jordan, C. H., Spencer, S. J., & Zanna, M. P. (2005). Types of high self-esteem and prejudice: How implicit self-esteem relates to ethnic discrimination among high explicit self-esteem individuals. *Personality and Social Psychology Bulletin*, *31*, 693–702.

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality & Social Psychology*, *85*(5), 969–978.

Katz, D. (1960). The functional approach to the study of attitudes. *Public Opinion Quarterly*, *24*(2), 163–204.

Kirkpatrick, L. A., Waugh, C. E., Valencia, A., & Webster, G. D. (2002). The functional domain specificity of self-esteem and the differential prediction of aggression. *Journal of Personality and Social Psychology*, *82*(5), 756–767.

Knobloch-Westerwick, S., & Meng, J. (2009). Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research*, *36*(3), 426–448.

Levin, K. D., Nichols, D. R., & Johnson, B. T. (2000). Involvement and persuasion: Attitude functions for the motivated processor. In G. R. Maio & J. M. Olson (Eds.), *Why we evaluate: Functions of attitudes* (pp. 163–194). Lawrence Erlbaum Associates.

Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs*, *49*(4), 800–823.

Luttrell, A., Petty, R. E., & Briñol, P. (2016). Ambivalence and certainty can interact to predict attitude stability over time. *Journal of Experimental Social Psychology*, *63*, 56–68.

Maio, G. R., & Haddock, G. (2004). Theories of attitudes: Creating a witches' brew. In G. Haddock & G. R. Maio (Eds.), *Contemporary perspectives on the psychology of attitudes* (pp. 425–453). Psychology Press.

Maio, G. R., & Haddock, G. (2015). *The psychology of attitudes and attitude change* (2nd ed.). SAGE.

Maio, G. R., & Olson, J. M. (2000a). Emergent themes and potential approaches to attitude function: The function-structure model of attitudes. In G. R. Maio & J. M. Olson (Eds.), *Why we evaluate: Functions of attitudes* (pp. 417–442). Lawrence Erlbaum Associates.

Maio, G. R., & Olson, J. M. (Eds.). (2000b). *Why we evaluate: Functions of attitudes*. Lawrence Erlbaum Associates.

Marsh, K. L., & Julka, D. L. (2000). A motivational approach to experimental tests of attitude functions theory. In G. R. Maio & J. M. Olson (Eds.), *Why we evaluate: Functions of attitudes* (pp. 271–294). Lawrence Erlbaum Associates.

McGregor, I., & Marigold, D. C. (2003). Defensive zeal and the uncertain self: What makes you so sure? *Journal of Personality & Social Psychology*, *85*(5), 838–852.

McGregor, I., Nail, P. R., Marigold, D. C., & Kang, S.-J. (2005). Defensive pride and consensus: Strength in imaginary numbers. *Journal of Personality & Social Psychology*, *89*(6), 978–996.

Miller, C. B. (2013). *Moral character: An empirical theory*. Oxford University Press.

Miller, C. B. (2014). *Character and moral psychology*. Oxford University Press.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268.

Olson, M. A., Fazio, R. H., & Hermann, A. D. (2007). Reporting tendencies underlie discrepancies between implicit and explicit measures of self-esteem. *Psychological Science*, *18*(4), 287–291.

Petrocelli, J. V., Tormala, Z. L., & Rucker, D. D. (2007). Unpacking attitude certainty: Attitude clarity and attitude correctness. *Journal of Personality and Social Psychology*, *92*(1), 30–41.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). Academic Press.

Petty, R. E., Wegener, D. T., & Fabrigar, L. R. (1997). Attitudes and attitude change. *Annual Review of Psychology*, *48*(1), 609–647.

Russell, D. C. (2009). *Practical intelligence and the virtues*. Oxford University Press.

Schröder-Abé, M., Rudolph, A., & Schütz, A. (2007). High implicit self-esteem is not necessarily advantageous: Discrepancies between explicit and implicit self-esteem and their relationship with anger expression and psychological health. *European Journal of Personality*, *21*(3), 319–339.

Sherman, S. J., & Fazio, R. H. (1983). Parallels between attitudes and traits as predictors of behavior. *Journal of Personality*, *51*(3), 308–345.

Snow, N. E. (2010). *Virtue as social intelligence: An empirically grounded theory*. Routledge.

Sosa, E. (2007). *A virtue epistemology: Apt belief and reflective knowledge* (Vol. 1). Clarendon Press.

Spencer, S. J., Jordan, C. H., Logel, C. E. R., & Zanna, M. P. (2005). Nagging doubts and a glimmer of hope: The role of implicit self-esteem in self-image maintenance. In A. Tesser, J. V. Wood, & D. E. Stapel (Eds.), *On building, defending, and regulating the self: A psychological perspective* (pp. 153–170). Psychology Press.

Tanesini, A. (2016a). I—"calm down, dear": Intellectual arrogance, silencing and ignorance. *Aristotelian Society Supplementary Volume*, *90*(1), 71–92.

Tanesini, A. (2016b). Teaching virtue: Changing attitudes. *Logos & Episteme*, *7*(4), 503–527.

Tanesini, A. (2018a). Intellectual humility as attitude. *Philosophy and Phenomenological Research*, *96*(2), 399–420.

Tanesini, A. (2018b). Epistemic vice and motivation [Special issue]. *Metaphilosophy*, *49*(3), 350–367.

Tanesini, A. (2018c). Intellectual servility and timidity. *Journal of Philosophical Research*, *43*, 21–41. https://doi.org/10.5840/jpr201872120

Thompson, M. M., & Zanna, M. P. (1995). The conflicted individual: Personality-based and domain specific antecedents of ambivalent social attitudes. *Journal of Personality*, *63*(2), 259–288.

Tormala, Z. L., & Rucker, D. D. (2007). Attitude certainty: A review of past findings and emerging perspectives. *Social and Personality Psychology Compass*, *1*(1), 469–492.

van Harreveld, F., van der Pligt, J., & de Liver, Y. N. (2009). The agony of ambivalence and ways to resolve it: Introducing the MAID model. *Personality and Social Psychology Review*, *13*(1), 45–61.

Vohs, K. D., & Heatherton, T. F. (2004). Ego threat elicits different social comparison processes among high and low self-esteem people: Implications for interpersonal perceptions. *Social Cognition*, *22*(1), 168–191.

Webber, J. (2006). Virtue, character and situation. *Journal of Moral Philosophy*, *3*(2), 193–213. https://doi.org/10.1177/1740468106065492

Webber, J. (2015). Character, attitude and disposition. *European Journal of Philosophy*, *23*(4), 1082–1096.

Zagzebski, L. T. (1996). *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge University Press.

Zagzebski, L. (2015). I-admiration and the admirable. *Aristotelian Society Supplementary Volume*, *89*(1), 205–221.

Zeigler-Hill, V., & Terry, C. (2007). Perfectionism and explicit self-esteem: The moderating role of implicit self-esteem. *Self and Identity*, *6*(2–3), 137–153.

Zunick, P. V., Teeny, J. D., & Fazio, R. H. (2017). Are some attitudes more self-defining than others? Assessing self-related attitude functions and their consequences. *Personality and Social Psychology Bulletin*, *43*(8), 1136–1149.

13

A Paradox of Information Aggregation

We Do It Well but Think About It Poorly, and Why This Is a Problem for Institutions

Hugo Mercier

How good are we at taking other people's opinion into account? Are we too easily swayed by authority figures or the majority opinion? Or, on the contrary, are we too pigheaded to change our minds even in the face of cogent arguments? In other words, how well do we take into account—*aggregate*—communicated information? (Throughout this chapter, I will refer to cases in which we take communicated information into account—by contrast with non-social sources of information—simply as *information aggregation*).

After having briefly exposed the great benefits that can be derived from information aggregation, I will make the case that there is a paradox in how humans deal with it: We are good at aggregating information but bad at thinking about it. More specifically, we are endowed with a suite of cognitive mechanisms that allow us to make the best of communicated information in a wide variety of contexts. These are naturally developing learning mechanisms of which we can already find roots in very young children. On the other hand, we are equipped with no such mechanisms when it comes to thinking about how information aggregation works or how well it works. Although we can learn—by reading chapters such as this one, say—about how information aggregation works, this remains explicit knowledge, distinct from the intuitive mechanisms with which we spontaneously aggregate information. Moreover, I argue the vast majority of people who do not learn such explicit knowledge rely instead on a variety of heuristics that happen to make them overly pessimistic about the prospects of information aggregation. To conclude, I delineate some of the consequences this pessimism might have for how we deal with modern information environments.

Hugo Mercier, *A Paradox of Information Aggregation* In: *Reason, Bias, and Inquiry.* Edited by: Nathan Ballantyne and David Dunning, Oxford University Press. © Oxford University Press 2022.
DOI: 10.1093/oso/9780197636916.003.0014

The Promise of Information Aggregation

That we can benefit from information aggregation is, in a sense, trivial: Other people know things we don't; more people are bound to know more things than any single individual. Yet the first mathematical demonstration of the power of information aggregation was only offered by Condorcet in the 18th century (Condorcet, 1785). What came to be known as the *Condorcet jury theorem* bears on the simple case of an assembly having to decide between two options, one of which is superior to the other. Condorcet demonstrated that the odds of the majority supporting the best option increase with the size of the assembly and that they converge to one for a sufficiently large assembly.

Because a series of conditions have to be met for the Condorcet jury theorem to hold, it seems to offer an upper bound on the power of information aggregation. The conditions are as follows. The choice has to be between two options, the assembly members must be at least minimally competent (i.e., have better than chance odds of selecting the correct option on their own), they must reveal their true preferences in their votes (i.e., no strategic voting), and they must have acquired their opinion independently of each other. However, subsequent mathematical analyses and modeling work have revealed that even when these assumptions are relaxed, the majority remains more likely than not to select the best option (Estlund, 1994; Feddersen & Pesendorfer, 1998; Hastie & Kameda, 2005; Ladha, 1992; Owen et al., 1989; Romeijn & Atkinson, 2011). The main result of the Condorcet jury theorem is quite robust.

Given its robustness, we can start thinking about the Condorcet jury theorem as being, in many cases, a lower bound on the power of information aggregation. The two main ways in which the Condorcet jury theorem offers only a lower bound to the power of information aggregation is that it grants each voice the same weight, regardless of its likelihood of being correct, and that it offers no opportunity to add new, potentially superior, options. Fortunately, in many cases of actual information aggregation, more weight can be put on some voices than on others—for instance, because we are able to track their past performance or to evaluate the reasons they provided to defend their opinion. This has the potential to greatly enhance the power of information aggregation. Likewise, the aggregation of information can allow the creation of new ideas and solutions that no individual could have thought of on their own (e.g., Page, 2007). It thus seems that information aggregation

has a huge potential to improve on individual decision-making (for review, see, e.g., Landemore, 2013).

We Are Good at Information Aggregation

That information aggregation has a lot of potential does not mean that people can make the best of it. In some cases, information aggregation cannot be performed optimally simply because people do not have the relevant information—they do not know which opinion is held by the majority, or they cannot learn of the reasons why different people hold different opinions. Here, we will look at cases in which the relevant information is available. This review focuses on work in experimental psychology, looking at two questions in turn: Are people able to take reliable cues into account in the way they aggregate information? Does this yield consistent improvements in their decision-making? (For in-depth reviews, see Mercier [2016, 2017, 2021]; Mercier & Morin [2021]; Mercier & Sperber [2011, 2017]; Sperber et al. [2010].)

Plausibility

People use their prior knowledge to evaluate the plausibility of communicated information. Information that is more coherent with our prior knowledge is more likely to be accepted (e.g., Bonaccio & Dalal, 2006; March et al., 2012; Petty & Cacioppo, 1979; Yaniv, 2004).

Majority

Do people know when they should follow majority opinion? On the whole, yes. People put more weight on the majority opinion when the majority is stronger and when the group to which the majority belongs is larger (R. Bond, 2005; Gerard et al., 1968; McElreath et al., 2005; Morgan et al., 2012). People are also able to discount majority opinions when the opinions have clearly not been acquired independently by the members of the majority (Mercier & Miton, 2021). However, there are also cases in which people do not weigh majority opinion properly. This is most likely to happen when

information is presented in a somewhat abstract format. For instance, knowing that a given percentage of a population supports an opinion has less effect than seeing a number of individuals holding the same opinion (Mutz, 1998). Similarly, knowing that several individuals tend to have highly correlated opinions doesn't lead to their opinion being discounted (Kroll et al., 1988; Maines, 1990).

Benevolence

One of the dangers inherent in the aggregation of social information lies in the conflicts of interests between senders and receivers. Through their communication, senders might seek to manipulate receivers in ways that serve senders' interests but not receivers'—for instance, they can lie. One reading of the literature on lie detection suggests that people pay attention to the wrong cues and cannot readily discern lies from honest statements (e.g., Global Deception Research Team, 2006). This pessimistic conclusion, however, largely stems from a combination of two factors. First is an over-reliance on explicit judgments about what cues people think are related to deception—for instance, averted gaze, which is not a reliable cue to deception. In fact, people do not so much rely on the cues they explicitly think are valid. Instead, they rely on more reliable cues, such as inconsistency or vagueness in the statements being evaluated (Hartwig & Bond, 2011). The second factor is that most lie-detection experiments require people to detect lies in the absence of any reliable cue to deception, such as background information (e.g., who has a motivation to lie). When such information is available, people make good use of it (C. F. Bond et al., 2013). On the whole, the experimental literature thus suggests that people are quite apt at detecting who is likely to attempt to deceive and to discount their statements appropriately. However, they do not do so using the cues on which most work has focused—behavioral cues—relying instead on more reliable cues, such as senders' incentives.

Confidence

In order to make sound individual decisions, people must be able to judge the degree of certainty with which they hold different beliefs (even if they do so using rough heuristics [Koriat, 2018]). It is thus likely that humans, along

with many cognitively complex animals, are endowed with mechanisms that track levels of confidence in their beliefs. Humans are arguably also endowed with mechanisms that allow them to communicate these degrees of confidence (Shea et al., 2014). To the extent that people are truthful, it would then make sense to take these communicated levels of confidence into account when aggregating information. Indeed, people seem to be able to communicate such degrees of confidence optimally (Bahrami et al., 2010). Moreover, people quickly become wary of speakers who abuse confidence signals by being consistently overconfident (Vullioud et al., 2016).

Competence

People can take a variety of cues to competence when deciding how much weight to put on an individual's opinion. They can look at past performance or at social markers of competence (e.g., believing a doctor's medical opinions) (e.g., Harvey & Fischer, 1997; Petty & Wegener, 1998). It has been suggested, however, that people tend to be too generous in their attribution of competence, following a kind of halo effect (Cooper, 1981) or prestige bias (Henrich & Gil-White, 2001). In either case, people would be overly deferential toward sources whose competence is suggested only by weak or irrelevant cues, from how admired someone is to how beautiful they are. Some experiments suggest that people might be weakly influenced by such cues but only when the decision is inconsequential and when more reliable cues are not available (e.g., Chaiken & Maheswaran, 1994). On the whole, people seem to appropriately restrict attributions of expertise to the proper domain of expertise (e.g., a doctor's medical opinion will be weighted more than their opinion in other domains). If anything, given the enormous differences in expertise we now encounter, we tend to underestimate the competence of experts, relative to ours, and thus to not put enough weight on their opinions (e.g., Motta et al., 2018).

Arguments

People are able to discriminate strong from weak arguments and to put more weight on opinions supported by the former than the latter. This has been observed using different measures of argument strength, whether they are

commonsensical (Petty & Wegener, 1998) or more formal (Bayesian formalism [Hahn & Oaksford, 2007]; argumentation theory [e.g., Hoeken et al., 2014, 2012; Hornikx, 2008]). There is no evidence that people are easily taken in by fallacious argument (such as egregious slippery slope arguments or arguments from ignorance [Corner et al., 2011; Oaksford & Hahn, 2004]). Moreover, people seem to change their minds when confronted with strong enough arguments, even when the arguments challenge confidently held opinions (Trouche et al., 2014, in press). Although some backfire effects have been reported (i.e., cases in which a good argument causes someone to shift their opinion away from the argument's conclusion [Nyhan & Reifler, 2010, 2015]), they seem to be rare, with good arguments moving their audience in the intended direction in most cases (e.g., Dockendorff & Mercier, 2021; Wood & Porter, 2019; for more references, see Mercier, 2021).

This very brief overview of the literature on the evaluation of communicated information shows that people are able to take a variety of cues into account in order to maximize the gains from information aggregation. Decisions should thus improve as a function of how many of these cues are available. Numerous pieces of evidence suggest that this is the case. Consider the results coming from advice-taking. This subfield of judgment and decision-making has examined how people take advice (i.e., other people's opinions) into account (for reviews, see Bonaccio & Dalal, 2006; Yaniv, 2004). When very little information is available about the source of the advice (e.g., only that it is another participant in the experiment), people do benefit from the advice but not as much as they could (Yaniv & Kleinberger, 2000). When more information is added—for instance, the degree of expertise of the source—people make better use of the advice (Harvey & Fischer, 1997). Similarly, being able to discuss with the source of the advice, and thus being exposed to their reasons for defending a given opinion, produces significant improvements over simply receiving their opinion (Liberman et al., 2012; Minson et al., 2011).

The same pattern emerges from the literature on forecasting. A line of effort to reach better forecasts started by simply averaging the forecasts of several experts. Forecast accuracy was improved by allowing the experts to see the opinions of other experts and to revise their opinions on this basis (Linstone & Turoff, 1976; for a more recent review, see Rowe & Wright, 1999). Forecast accuracy was improved again when experts had access to the reasons behind the other experts' opinions (Rowe & Wright, 1996) or when they could directly discuss with each other (Mellers et al., 2014).

The context in which the most cues relevant for information aggregation are available is that of a group discussion, especially a group discussion taking place between people who know each other (Michaelsen et al., 1989). Group discussion allows group members to express their confidence, demonstrate their expertise, and spell out the reasons behind their opinions. These reasons can be debated, bad reasons shot down, and good reasons accepted and elaborated on. As a result, the most reliable cues tend to win. For instance, when clear reasons can be provided for the best answer to a problem, these reasons carry the day, even in the face of other cues, such as majority or confidence (Trouche et al., 2014). When reasons do not have such a perfect discriminating power, they still play a role; but they are balanced out by other cues, such as majority (Laughlin, 2011). This explains why group discussion leads to consistent improvements in performance in a wide range of domains (for reviews, see Mercier, 2016; Mercier & Sperber, 2011), as long as some minimal conditions are met (group members must share some overarching goal, and there must be some disagreement within the group).

We Are Bad at Thinking About Information Aggregation

The evidence reviewed so far shows that people take a variety of cues into account when aggregating information—cues that pertain to the content of the communicated information as well as to its source(s). More specifically, the evidence suggests that people are equipped with a set of specialized cognitive mechanisms that evaluate these cues but not in a way that requires conscious awareness of which cues are taken into account and why they are taken into account (with a partial exception for arguments). For instance, in the case of lie detection, people mostly take into account reliable cues to deception; but when asked which cues to deception are reliable, they indicate unreliable cues (e.g., gaze aversion). Here, I will briefly review some evidence suggesting that people do not have accurate explicit beliefs about the advantages of information aggregation.

I have mentioned some studies suggesting that when people see that many others hold the same opinion, they are more likely to accept it. Other studies, however, suggest that when information about majority opinion is presented in a more abstract way (e.g., 90% of people surveyed agree with X), it is much less effective (Mutz, 1998). A recent study examined more precisely people's grasp of the main result from the Condorcet jury theorem, revealing that

it was completely lacking (Mercier et al., 2021). For instance, in one of the studies, participants had to evaluate the odds that majority voting would lead an assembly to select the best of two options. The conditions specified ensured that the Condorcet jury theorem would apply (in particular, assembly members were described as competent). Yet the average answer for the odds that the assembly would select the best option were approximately equal to the odds that a single member would select the best option. People had no intuition whatsoever that aggregating votes would lead to an improvement in the odds of selecting the best option.

Similar results have been previously obtained in relation with averaging. When presented with two numerical opinions (e.g., forecasts) and no strong cue that one is superior to the other, averaging between the two is much more likely to yield an accurate opinion than choosing one of the two opinions—the so-called averaging principle (Larrick & Soll, 2006). However, when people are explicitly asked about this or when they have to base their opinion on two such opinions (with no other cue available), they display no understanding of the averaging principle (Larrick & Soll, 2006; Mercier et al., 2012; Soll & Larrick, 2009).

We observe the same pattern when people are asked about the efficacy of group discussion. In a series of experiments, participants from various backgrounds (including experts in the task at hand) were presented with a reasoning task, asked to solve it (for all participants but the experts), then asked to estimate the odds that a single individual, and then a small group discussing together, would solve the task (Mercier et al., 2015). Well-established results show that, on this task, groups outperform individuals by a factor of 5 on average (going from 12% correct answers to 60%). However, most participants estimated that group discussion would bring no benefit. Even experts significantly underestimated the effect of group discussion. For instance, they thought that in a discussion someone with the correct answer would only convince someone with the wrong answer in 70% of cases, when the real number is close to 100%. Although such drastic improvements are mostly observed with tasks that have an accessible, demonstrative answer, group discussion improves performance in a wide range of domains. Given that it's unclear why people would become more optimistic about the outcome of group discussion in other domains, we can surmise that the underestimation of the value of group discussion is quite robust.

More generally, people seem to take a dim view of other people's ability to aggregate information. For instance, people do not think they are overly

influenced by the media, but they think other people are (the so-called third-party media effect; for review, see, Mutz [1998]).

These findings are mirrored in the historical record. Many scholars have taken—and keep taking—a very dim view of most people's ability to aggregate information. The idea that people gullibly defer to prestigious figures, irrespective of their domain of expertise, or that they blindly follow majority opinions is quite prevalent. This quotation from a recent book by a political philosopher is somewhat extreme but nonetheless telling:

> Actual human beings are wired not to seek truth and justice but to seek consensus. They are shackled by social pressure. They are overly deferential to authority. They cower before uniform opinion. They are swayed not so much by reason but by a desire to belong, by emotional appeal, and by sex appeal. (Brennan, 2012, p. 8)[1]

We find similar ideas in the work of prominent psychologists ("That human beings are, in fact, more gullible than they are suspicious should probably 'be counted among the first and most common notions that are innate in us' " [Gilbert et al., 1990, p. 231]) and other scholars (see references in Mercier [2017]). Arguably, the idea that people gullibly follow leaders has had a significant historical impact. For instance, Jason Stanley has argued that fear of demagogues has been "political philosophy's central reason for skepticism about democracy" (2015, p. 27).

That scholars have criticized our supposed tendency to blindly follow the majority might be related to the dim view they often take of the power of information aggregation to yield felicitous outcomes. Nineteenth-century crowd psychologists were among the worst offenders in this respect, rejecting strongly all attempts at collective deliberation, whether they be congresses, parliaments, or juries (Barrows, 1981). Indeed, they even defended the idea that group deliberation made matters worse ("In any case, it is clear that the jury is even less intelligent than the jurors" [Tarde, 1895, p. 23, my translation]).

Fortunately, we also find some exceptions to this common denigration of the power of information aggregation. The most historically significant is a passage from the third book of Aristotle's *Politics*, in which he offers a defense of the wisdom of the crowd (e.g. "The many, of whom none is individually

[1] I thank Olivier Morin for pointing this quote out to me.

an excellent man, nevertheless can when joined together be better—not as individuals but all together—than those [who are best]" [2013, III, 11]). It seems that Aristotle's arguments did not find much of an echo until the Enlightenment, for example, with the discovery of the jury theorem by Condorcet (although it was then forgotten until the mid-20th century [see Dietrich & Spiekermann, 2013]). A similar sentiment was then found in defense of freedom of speech, in what would become "marketplace of ideas" arguments (e.g., Mill, 1974; Thomas Jefferson was another famous proponent of this argument and John Milton an early one).

In spite of these famous exceptions, I would argue that defending the power of information aggregation has been, historically, a minority position. Instead, a great many scholars have held dim views of the potential of information aggregation and of people's ability to make use of whatever potential there might be. There is thus a sharp contrast between how good people are at information aggregation and how bad they are at thinking about it. The contrast should not be particularly surprising. Throughout our recent evolution, there must have been significant selective pressures on our abilities to aggregate information (Mercier & Sperber, 2017; Sperber et al., 2010). By contrast, an explicit understanding of the principles of information aggregation would not have been of much use. Indeed, we find similar contrasts in every domain of cognition. For instance, we are equipped with specialized mechanisms that approximate some laws of physics, allowing us to move about and interact with objects. However, any explicit belief we might have about physics—our naïve physics—is hopelessly out of touch.[2]

Information Formats

The fact that we have no explicit grasp of the principles of physics usually does not stop us from interacting properly with the objects that surround us. Most of us interact with these objects in ways that are not too different from the ways in which our ancestors interacted with the objects in their environment (from a physics point of view). Clearly, there are exceptions—flying a

[2] What might be surprising, then, is not that people easily adopt inaccurate explicit beliefs regarding information aggregation but that these inaccurate beliefs tend to veer in the direction of pessimism regarding the power of information aggregation. Why aren't people just as likely to be overly optimistic? This systematic bias likely stems from a conjunction of factors, which I will not explore here (but see Mercier [2017] for some suggestions regarding the widespread belief that people are gullible).

plane or building a skyscraper, say—but for most of us in daily life, there is no need for accurate, explicit physical theories. This might not be the case when it comes to information aggregation, for our informational environment, compared to the one faced by our ancestors, has been dramatically modified and expanded.

One of the many differences in our information environment—compared to the environment we evolved in—is that information reaches us after having passed through a great many more intermediaries. In a small-scale society, the vast majority of communicated information (by contrast with technical skills) would be second-hand, sometimes third-hand, but it would rarely have gone through more intermediaries (traditions such as those pertaining to the supernatural being one of these exceptions [see Boyer, 2001; Morin, 2015]). Nowadays, we acquire a significant amount of information through the media (either directly or through friends, colleagues, etc. [see Lazarsfeld et al., 1948]). Take a scientific finding: It already requires many intermediaries before a scientific article is published. It then goes through a press release, an article in a newspaper, then maybe a colleague before it reaches you. One of the consequences of the introduction of these many intermediaries is that we only have indirect information about the ultimate source(s) of the information. We have not directly witnessed their competence or had opportunities to evaluate their trustworthiness. We haven't talked to a large number of sources to get an intuitive sense of the level of consensus. As a result, when we are presented with the information, we are much less well equipped to judge its accuracy. This will often mean that we are overly skeptical of information that has gone through multiple intermediaries since the original reasons for accepting the information will be lost in the process of transmission.

Take information about climate change. Several experiments have attempted to convince participants of the existence of anthropogenic climate change by presenting them with information about the high degree of consensus among climate scientists (e.g., Lewandowsky et al., 2013; van der Linden et al., 2015). Results have been somewhat contradictory, but on the whole, it seems that such information has only a moderate impact on participants' beliefs (Kahan, 2017). One can easily imagine that, by contrast, witnessing first-hand the degree of consensus among such a large number of scientists, accompanied by some first-hand knowledge of their competence and trustworthiness, would be vastly more convincing. Not only does one have more information in the case, but this information is also presented

in a format that makes it more intuitively compelling. A better abstract understanding of the principles of information aggregation or a better understanding of how science is conducted might help close this gap. People would then be better able to properly evaluate some of the information that has been condensed as it went through intermediaries.

Explicit Beliefs About Information Aggregation in Institutions

Another major difference between our current environment and that of our ancestors is that we now often find ourselves in institutionally mandated information environments. Students, jurors, elected representatives, and many others all find themselves in environments in which the people they interact with, and the ways they interact with these people, are largely determined by an institutional framework and not by their own choices. For instance, in most cases students are not allowed to talk to each other in class or to collaborate when facing a test. This constrains how people can aggregate information.

These complex institutions are not simply designed according to any creator's grand plans. Instead, they evolve willy-nilly, the result of a wide array of forces. However, in some cases at least, some people play a prominent role in molding these institutions—members of constitutional conventions or school boards, say. Those individuals who play an important role in molding institutions have a variety of incentives. Some of these incentives are problematic because they do not fit with those of the people who will be subject to the institution. Undoubtedly, this type of "principal-agent" problem might yield suboptimal institutional designs (at least from the point of view of those subject to the institution). For instance, some pedagogical methods appear chiefly aimed at teaching to the test, which might be better for the people running administrations than for the students.

Fortunately, sometimes people who mold institutions have an incentive to maximize the institution's capacity to aggregate information. One would hope that people who can influence how schools work care about how children learn, that people who decide how juries make decisions care about whether they deliver the right verdict, and so on. At this stage, explicit beliefs about information aggregation might play a role. The realization that information aggregation yields significant epistemic and practical benefits and

that most individuals are apt at making the best of these benefits under relatively simple conditions (e.g., group discussion) should lead one to design institutions accordingly. Unfortunately, as I've argued, this realization has been, historically, rather rare. As a result, many institutions might have suboptimal features when it comes to their capacity to aggregate information.

In what follows, I will use group discussion as an example since it is a context in which information aggregation proves very efficient under a wide range of circumstances and since group discussion is, in most cases, very easy to implement (by contrast with, say, prediction markets). Obviously, people discuss and exchange arguments with each other in just about every institution (monastic orders enforcing vows of silence being the rare exception). But what I'm interested in are formal, institutionally mandated forms of group discussion, such as jury deliberation. I will attempt to show that few institutions mandating group discussion do so on the basis of an explicit belief in the virtues of information aggregation. The main exception will be cases in which these explicit beliefs have been buttressed by empirical evidence.

Parliaments

In a loose sense of the word—an assembly representing the members of a given society that deliberates to make collective decisions—parliaments are very old indeed. For example, most European societies had some similar form of assembly in the first half of the first millennium, and some persisted for a long time (most famously the Nordic *thing*). These assemblies can be seen as the descendants of even older forms of collective decision-making, as can be found in hunter–gatherer societies, that rely on group discussion to reach important decisions (Boehm et al., 1996). In both cases—even if this is clearer in the case of hunter–gatherers—these assemblies are necessary because no single individual holds enough power to force a collective decision. A significant part of the population (e.g., adult males) has to be made to agree to the decision before it can be in any way implemented. As a result, there is no need for anyone to realize that such procedures often lead to superior decisions.

If we turn to more formal assemblies, such as the Roman Senate, it is likely that some of their most important rules were the outcome of power struggles rather than concerns about the assembly's efficacy. For instance,

supermajority rules appeared in order to stop Roman senators from being too easily condemned (Schwartzberg, 2014). If we now turn to recent and contemporary parliamentary procedures, we observe a similar pattern. These procedures seem to be largely governed by the short-term political interests of the agents in charge, rather than by concerns about whether the procedures would allow the parliament to deliberate more efficiently (Binder, 1997; Cox & McCubbins, 1993; Dion, 2001). On the whole, it is thus very unlikely that an acknowledgment of the power of information aggregation played a significant role in shaping parliaments and parliamentary procedures.

Juries

From an Anglo-Saxon perspective, juries and jury deliberation might seem like an essential part of the judicial system. Moreover, the arguments raised in defense of juries suggest a clear understanding of the epistemic advantages of the information aggregation taking place during jury deliberation (see, e.g., Ellsworth, 1989; Hastie et al., 1983). It thus seems that jury deliberation might be a case in which a positive view of information aggregation helped spread an institution that makes good use of group discussion. However, one should keep in mind that juries are a historical anomaly. For several centuries after their appearance, they could only be found in common-law countries (most famously England). Juries seem to have been introduced not necessarily to make better judicial decisions but to make more decisions that appeared more legitimate: They replaced the ordeal as a way of justifying capital punishment (Fisher, 1997) (on the continent, confessions and the torture used to extract them played the same role [Langbein, 2012]).

The extension of juries to civil-law countries has been haphazard and is, in some cases, very recent (e.g., the jury was only introduced in Japan in the early 2000s [see Anderson & Ambler, 2006]). Moreover, juries are still resisted, not only in civil-law countries (e.g., France and Belgium [see Frydman, 2007]) but also in common-law countries. For instance, in the United States prominent scholars have mounted a systematic—but deeply flawed (Feigenson, 2003; Vidmar, 2004)—attack on the reliance of juries to award punitive damages (Sunstein, 2002). This latter attack is symptomatic of a dim view of the power of information aggregation since it specifically suggests that group discussion systematically leads to worse decisions (Schkade et al., 2000).

I would thus argue that juries owe little of their cultural success to a positive view of their epistemic value.

Schools

Group discussion can play at least two roles in school: Group discussion between teachers can help improve their teaching skills (e.g., through lesson study [see Fernandez, 2002]), and group discussion between students can help them learn better (through collaborative or cooperative learning). To the best of my knowledge, few school systems mandate discussion between teachers in order to improve their teaching (in spite of mounting evidence in favor of such methods [see Ming Cheung & Yee Wong, 2014), so I will focus on collaborative learning between students. Prominent scholars have emphasized the importance of sociality for learning, but most have focused on the relation between student and teacher—as in the Socratic method or the work of Vygotsky (e.g., 1978). The idea that group discussion between students might be beneficial only became popular relatively recently (Bruffee, 1984). Interestingly, the inception of collaborative learning might owe more to political ideals—anti-authoritarianism in particular—than to beliefs about the epistemic value of discussion per se (Bruffee, 1984). The subsequent spread of collaborative learning, in particular in the United Kingdom and the United States, was propelled by the accumulating, and by now massive, evidence showing its potential to improve learning outcomes (for reviews, see, e.g., Slavin, 1996, 2014). Yet the use of collaborative learning remains rare in many countries, even countries with high investment in education (Algan et al., 2013). There might be many obstacles explaining the so far limited spread of collaborative learning—teachers might find it difficult to implement, for instance. Even so, what the history of collaborative learning suggests is that its spread has been largely driven by the evidence gathered in its favor and not by a pre-existing belief in the power of student discussion to yield better epistemic outcomes.

Science

Clearly, discussion plays a crucial role in science. In many cases, these discussions are institutionally mandated—from peer review to conferences

or even lab meetings. Moreover, some of the thinkers who were at the forefront of the scientific revolution of the 17th century recognized the epistemic value of discussion—Robert Boyle being the best-known example (see Shapin & Schaffer, 1985). It might thus seem that science is the one counterexample in which the recognition of the advantage of information aggregation—prior to any real evidence in its support—led to the institutional implementation of group discussions. However, other, equally influential thinkers had little time for group discussion (e.g., Bacon, 1620).

More importantly, many institutions that mandate group discussion in science likely emerged simply out of the need and desire of scientists to engage in group discussion. Scientists are typically eager to share their findings. In so doing, they can barely avoid critical discussions—indeed, they are generally keen on criticizing others' theories and results. Explicit beliefs about the virtues of information aggregation are not necessary. Moreover, some forms of discussion that have become paramount in science clearly did not emerge because of an explicit belief in their epistemic virtues. Most prominently, peer review—which has become one of the defining traits of science—emerged as a form of censorship (Biagioli, 2002). Scientists started reviewing each other's manuscripts in order to preempt harsher royal censorship. The idea that this form of discussion would improve the scientific content of the texts was thus largely absent from peer review's origins, and it is not clear whether it has played a significant role in peer review's persistence (although see, Csiszar, 2016).

This very brief overview suggests that institutions mandating group discussion can appear for a variety of reasons. In some cases, individuals simply want to take part in discussions, and they find ways of formalizing these interactions (science being the prominent example). In other cases, institutionally mandated group discussion is the outcome of power struggles (on the role of power in shaping institutions, see, e.g., Knight [1995]). To some extent, parliaments and juries are a concession extracted from powerful agents by a broader community. We also find a hodgepodge of specific rationales—for instance, the anti-authoritarian strain in the early proponents of collaborative learning—but on the whole an explicit recognition of the potential epistemic and practical benefits of group discussion does not seem to have played a prominent role in the development of the institutions that rely on group discussion.

The only exception appears when there is clear feedback on the efficacy of group discussion. This feedback can be relatively informal—some judges

might have been able to appreciate first-hand the benefits of jury deliberation—or it can be very formal—the scientific evidence accumulated in favor of collaborative learning. To the extent that we do not spontaneously form accurate explicit beliefs about information aggregation, this makes sense. The fact that such feedback is hard to come by—in some cases, it takes randomized control trials for the evidence to clearly emerge—would then help explain the underuse of group discussion in many institutional contexts. Here, the contrast with naïve physics is telling. Any attempt to build a complex structure provides immediate feedback: Our attempt is successful or not. It is thus easy to realize that our naïve beliefs about physics are largely misguided. By contrast, it is difficult to judge the outcome of a complex institution and even more difficult to understand what individual factors are responsible for the outcome.

Popularizing the Wisdom of Crowds

I have argued that people are endowed with a set of cognitive mechanisms that allow them to efficiently aggregate information in a range of contexts. Unfortunately, people increasingly function within institutional frameworks that put stringent limits on their ability to aggregate information (e.g., limiting students' ability to work in groups). Many of our institutions are thus far from making an optimal use of our abilities to aggregate information. This failure stems from a combination of many factors. One factor might be the apparent difficulty in grasping the power of information aggregation and the widespread pessimism regarding others' ability to make good use of information aggregation.

In recent years, mounting evidence has started to buck the trend. In some countries at least, schools rely increasingly more on collaborative learning. We find analogous moves in other domains, for instance, the increased use of teams in medical decision-making and forecasting. We might even hope that this evidence-based optimism toward the power of information aggregation spreads further. The popular success of books such as Surowiecki's *The Wisdom of Crowds* (2005) and Philip Tetlock and Dan Gardner's *Superforecasting: The Art and Science of Prediction* (2016) can help spread much needed optimism. We can then only hope that once they are widespread, these more positive, but also more accurate, beliefs about the power

of information aggregation will help push for institutional designs that make a better use of our abilities to aggregate information.

Acknowledgments

I thank Pascal Boyer, Christophe Heintz, Adrien Aufort, and the editors for their precious feedback. This work was supported by the Agence Nationale de la Recherche grant EUR FrontCog ANR-17-EURE-0017 and ANR-10-IDEX-0001-02 to PSL, and by the grant "An Evolutionary and Cultural Perspective on Intellectual Humility via Intellectual Curiosity and Epistemic Deference" from the John Templeton Foundation.

References

Algan, Y., Cahuc, P., & Shleifer, A. (2013). Teaching practices and social capital. *American Economic Journal: Applied Economics*, *5*(3), 189–210.

Anderson, K., & Ambler, L. (2006). The slow birth of Japan's quasi-jury system (*saiban-in seido*): Interim report on the road to commencement. *Journal of Japanese Lax*, *11*, 55–80.

Aristotle. (2013). *Aristotle's "Politics"* (C. Lord, Trans.). University of Chicago Press.

Bacon, F. (1853). *Novum organum*. In The Physical and Metaphysical Works of Lord Bacon, book 1. London: H. G. Bohn.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085.

Barrows, S. (1981). *Distorting mirrors: Visions of the crowd in late nineteenth-century France*. Yale University Press.

Biagioli, M. (2002). From book censorship to academic peer review. *Emergences: Journal for the Study of Media & Composite Cultures*, *12*(1), 11–45.

Binder, S. A. (1997). *Minority rights, majority rule: Partisanship and the development of Congress*. Cambridge University Press.

Boehm, C., Antweiler, C., Eibl-Eibesfeldt, I., Kent, S., Knauft, B. M., Mithen, S., Richerson, P J., & Wilson, D. S. (1996). Emergency decisions, cultural-selection mechanics, and group selection [and comments and reply]. *Current Anthropology*, *37*(5), 763–793.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151.

Bond, C. F., Howard, A. R., Hutchison, J. L., & Masip, J. (2013). Overlooking the obvious: Incentives to lie. *Basic and Applied Social Psychology*, *35*(2), 212–221.

Bond, R. (2005). Group size and conformity. *Group Processes & Intergroup Relations*, *8*(4), 331–354.

Boyer, P. (2001). *Religion explained*. Heinemann.

Brennan, J. (2012). *The ethics of voting*. Princeton University Press.

Bruffee, K. A. (1984). Collaborative learning and the "conversation of mankind." *College English*, *46*(7), 635–652.

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology*, *66*(3), 460–473.

Condorcet. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'Imprimerie royale.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, *90*(2), 218–244.

Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, *64*(2), 133–152.

Cox, G. W., & McCubbins, M. D. (1993). *Legislative leviathan: Party government in the House* (California Series on Social Choice and Political Economy 23). University of California Press.

Csiszar, A. (2016). Troubled from the start: Pivotal moments in the history of academic refereeing have occurred at times when the public status of science was being renegotiated. *Nature*, *532*(7599), 306–309.

Dietrich, F., & Spiekermann, K. (2013). Epistemic democracy with defensible premises. *Economics and Philosophy*, *29*(1), 87–120.

Dion, D. (2001). *Turning the legislative thumbscrew: Minority rights and procedural change in legislative politics*. University of Michigan Press.

Dockendorff, M., & Mercier, H. (2021). *Argument transmission as the weak link in the correction of political misbeliefs* [Unpublished manuscript].

Ellsworth, P. C. (1989). Are twelve heads better than one? *Law and Contemporary Problems*, *52*, 205–224.

Estlund, D. (1994). Opinion leaders, independence, and Condorcet's jury theorem. *Theory and Decision*, *36*(2), 131–162.

Feddersen, T., & Pesendorfer, W. (1998). Convicting the innocent: The inferiority of unanimous jury verdicts under strategic voting. *American Political Science Review*, *92*(1), 23–35.

Feigenson, N. R. (2003). Can tort juries punish competently [Book review]. *Chicago-Kent Law Review*, *78*(1), 239.

Fernandez, C. (2002). Learning from Japanese approaches to professional development: The case of lesson study. *Journal of Teacher Education*, *53*(5), 393–405.

Fisher, G. (1997). The jury's rise as lie detector. *The Yale Law Journal*, *107*(3), 575–713.

Frydman, B. (2007). La contestation du jury populaire. Symptôme d'une crise rhétorique et démocratique. *Questions de Communication*, *12*, 103–117.

Gerard, H. B., Wilhelmy, R. A., & Conolley, E. S. (1968). Conformity and group size. *Journal of Personality and Social Psychology*, *8*(1, Pt 1), 79–82.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601–613.

Global Deception Research Team. (2006). A world of lies. *Journal of Cross-Cultural Psychology*, *37*(1), 60–74.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732.

Hartwig, M., & Bond, C. H. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*, 117–133.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*(2), 494–508.

Hastie, R., Penrod, S., & Pennington, N. (1983). *Inside the jury*. Harvard University Press.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, *22*(3), 165–196.

Hoeken, H., Šorm, E., & Schellens, P. J. (2014). Arguing about the likelihood of consequences: Laypeople's criteria to distinguish strong arguments from weak ones. *Thinking & Reasoning*, *20*(1), 77–98.

Hoeken, H., Timmers, R., & Schellens, P. J. (2012). Arguing about desirable consequences: What constitutes a convincing argument? *Thinking & Reasoning*, *18*(3), 394–416.

Hornikx, J. (2008). Comparing the actual and expected persuasiveness of evidence types: How good are lay people at selecting persuasive evidence? *Argumentation*, *22*(4), 555–569.

Kahan, D. (2017). The "gateway belief" illusion: Reanalyzing the results of a scientific-consensus messaging study. *Journal of Science Communication*, *16*(5), 1–20.

Knight, J. (1995). Models, interpretations, and theories: Constructing explanations of institutional emergence and change. In J. Knight & I. Sened (Eds.), *Explaining social institutions* (pp. 95–120). University of Michigan Press.

Koriat, A. (2018). When reality is out of focus: Can people tell whether their beliefs and judgments are correct or wrong? *Journal of Experimental Psychology: General*, *147*(5), 613–631.

Kroll, Y., Levy, H., & Rapoport, A. (1988). Experimental tests of the separation theorem and the capital asset pricing model. *The American Economic Review*, *78*(3), 500–519.

Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, *36*(3), 617–634.

Landemore, H. (2013). Democratic reason: The mechanisms of collective intelligence in politics. In J. Elster & H. Landemore (Eds.), *Collective wisdom* (pp. 251–289). Cambridge University Press.

Langbein, J. H. (2012). *Torture and the law of proof: Europe and England in the Ancien Régime*. University of Chicago Press.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*, 111–127.

Laughlin, P. R. (2011). *Group problem solving*. Princeton University Press.

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1948). *The people's choice: How the voter makes up his mind in a presidential campaign*. Columbia University Press.

Lewandowsky, S., Gignac, G. E., & Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change*, *3*(4), 399–404.

Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the "wisdom of dyads." *Journal of Experimental Social Psychology*, *48*(2), 507–512.

Linstone, H. A., & Turoff, M. (Eds.). (1976). *The Delphi method: Techniques and applications* (Vol. 18). Addison-Wesley.

Maines, L. A. (1990). The effect of forecast redundancy on judgments of a consensus forecast's expected accuracy. *Journal of Accounting Research*, *28*, 29–47.

March, C., Krügel, S., & Ziegelmeyer, A. (2012). *Do we follow private information when we should? Laboratory evidence on naıve herding* [Jena Economic Research Paper 002]. Friedrich-Schiller-University Jena.

McElreath, R., Lubell, M., Richerson, P. J., Waring, T. M., Baum, W., Edsten, E., Efferson, C., & Paciotti, B. (2005). Applying evolutionary models to the laboratory study of social learning. *Evolution and Human Behavior*, *26*(6), 483–508.

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.

Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700.

Mercier, H. (2017). How gullible are we? A review of the evidence from psychology and social science. *Review of General Psychology*, *21*(2), 103–122.

Mercier, H. (2020). *Not Born Yesterday: The Science of Who we Trust and What we Believe*. Princeton University Press.

Mercier, H., Dockendorff, M., Majima, Y., Hacquin, A.-S., & Schwartzberg, M. (2021). Intuitions about the epistemic virtues of majority voting. *Thinking & Reasoning*, *27*(3), 445–463.

Mercier, H., & Miton, H. (2019). Utilizing simple cues to informational dependency. *Evolution and Human Behavior*, *40*(3), 301–314.

Mercier, H., & Morin, O. (2019). Majority rules: How good are we at aggregating convergent opinions? *Evolutionary Human Sciences*, *1*, e6.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

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

Mercier, H., Trouche, E., Yama, H., Heintz, C., & Girotto, V. (2015). Experts and laymen grossly underestimate the benefits of argumentation for reasoning. *Thinking & Reasoning*, *21*(3), 341–355.

Mercier, H., Yama, H., Kawasaki, Y., Adachi, K., & Van der Henst, J.-B. (2012). Is the use of averaging in advice taking modulated by culture? *Journal of Cognition and Culture*, *12*(1–2), 1–16.

Michaelsen, L. K., Watson, W. E., & Black, R. H. (1989). A realistic test of individual versus group consensus decision making. *Journal of Applied Psychology*, *74*(5), 834–839.

Mill, J. S. (1974). *On liberty*. Pelican Books.

Ming Cheung, W., & Yee Wong, W. (2014). Does lesson study work? A systematic review on the effects of lesson study and learning study on teachers and students. *International Journal for Lesson and Learning Studies*, *3*(2), 137–149.

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to tango. *Personality and Social Psychology Bulletin*, *37*(10), 1325–1338.

Morgan, T. J. H., Rendell, L. E., Ehn, M., Hoppitt, W., & Laland, K. N. (2012). The evolutionary basis of human social learning. *Proceedings of the Royal Society of London B: Biological Sciences*, *279*(1729), 653–662.

Morin, O. (2015). *How traditions live and die*. Oxford University Press.

Motta, M., Callaghan, T., & Sylvester, S. (2018). Knowing less but presuming more: Dunning-Kruger effects and the endorsement of anti-vaccine policy attitudes. *Social Science & Medicine*, *211*, 274–281.

Mutz, D. C. (1998). *Impersonal influence: How perceptions of mass collectives affect political attitudes*. Cambridge University Press.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine*, *33*(3), 459–464.

Oaksford, M., & Hahn, U. (2004). A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, *58*(2), 75–85.

Owen, G., Grofman, B., & Feld, S. L. (1989). Proving a distribution-free generalization of the Condorcet jury theorem. *Mathematical Social Sciences*, *17*(1), 1–16.

Page, S. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press.

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology*, *37*, 349–360.

Petty, R. E., & Wegener, D. T. (1998). Attitude change: Multiple roles for persuasion variables. In D. T. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 323–390). McGraw-Hill.

Romeijn, J., & Atkinson, D. (2011). A Condorcet jury theorem for unknown juror competence. *Politics, Philosophy, and Economics*, *10*(3), 237–262.

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, *12*(1), 73–89.

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting*, *15*(4), 353–375.

Schkade, D., Sunstein, C. R., & Kahneman, D. (2000). Deliberating about dollars: The severity shift. *Columbia Law Review*, *100*, 1139–1176.

Schwartzberg, M. (2014). *Counting the many: The origins and limits of supermajority rule*. Cambridge University Press.

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton University Press.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193.

Slavin, R. E. (1996). Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology*, *21*(1), 43–69.

Slavin, R. E. (2014). Cooperative learning and academic achievement: Why does groupwork work? *Anales de Psicología/Annals of Psychology*, *30*(3), 785–791.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, *25*(4), 359–393.

Stanley, J. (2015). *How propaganda works*. Princeton University Press.

Sunstein, C. R. (2002). *Punitive damages: How juries decide*. University of Chicago Press.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books.

Tarde, G. (1895). *Essais et mélanges sociologiques*. A. Storck.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.

Trouche, E., Shao, J., & Mercier, H. (2019) Objective evaluation of demonstrative arguments. *Argumentation*, *33*(1), 23–43.

van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PLoS One*, *10*(2), Article e0118489.

Vidmar, N. (2004). Experimental simulations and tort reform: Avoidance, error, and overreaching in Sunstein et al.'s punitive damages. *Emory Law Journal*, *53*, 1359–1403.

Vullioud, C., Clément, F., Scott-Phillips, T., & Mercier, H. (2016). Confidence as an expression of commitment: Why misplaced expressions of confidence backfire. *Evolution and Human Behavior*, *38*(1), 9–17.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*, 135–163.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*, 1–13.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*, 260–281.

14

Why Do People Argue Past One Another Rather than with One Another?

Deanna Kuhn and Kalypso Iordanou

Both educators and employers emphasize the importance of proficient thinking and learning in today's rapidly changing contexts, more than one's accumulated knowledge. Yet people don't exchange ideas all that well. Serious public discourse is at a disturbingly low level in contemporary American society, largely confined to echo chambers dominated by sound bites and slogans (Barbera et al., 2015).

The values of contemporary culture offer one explanation for this state of affairs. Here, we consider whether there also exist more enduring factors than current cultural ones, factors at the level of the individual. One of several possibilities is the dominance of emotion. T. S. Eliot wrote, "[W]hen we do not know, or when we do not know enough, we tend always to substitute emotions for thoughts" (Eliot, 1921/2013). Recent research shows that high affect is not associated with arguments that are any stronger; it only enhances the belief that an argument is more persuasive, what has been called the "illusion of argument justification" (Fisher & Keil, 2014).

A second possibility is socially driven. We are motivated to protect our beliefs, not exposing them to the views of others who may challenge them. So we don't listen unless we already know we agree. "Tell me something I already know" is a regular request put to guests by host Jordan Klepper of Comedy Central's *The Opposition*. In a more serious vein, in response to a reporter's query, "Do you think that talking about millions of illegal votes is dangerous to this country without presenting the evidence?" the 45th US president said, "No. Not at all. Not at all, because many people feel the same way I do."[1] Agreement thus eliminates the need for any further standards of verification.

[1] Asked by David Muir, the anchor of ABC's *World News Tonight*, of Donald Trump in January 2017 (quoted in Andersen, 2017).

Deanna Kuhn and Kalypso Iordanou, *Why Do People Argue Past One Another Rather than with One Another?* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press. DOI: 10.1093/oso/9780197636916.003.0015

A third, recently posed possibility is evolutionary. Humans have developed skills of argument not for debating one another but as a tool for promoting individual views in social contexts and thereby advancing personal objectives (Mercier & Sperber, 2011). We do not seek to sharpen these views by means of intellectual exchange, nor is argumentation necessarily aimed at truth.

These three possibilities at the level of the individual—emotional, motivational, and evolutionary—are not mutually exclusive and are most likely additive or interactive. The second and third ones, for example, are consistent with an individual's concealing contradictory arguments or evidence for strategic advantage.

In the remainder of this chapter, we focus on a fourth potential individual factor, one that is likely to interact with the others: an individual's reasoning capabilities. Such capabilities, we know, show an identifiable course of development during the first decade or two of the human life span and wide individual differences thereafter. Could these differences impose limitations on the effectiveness of the discussion and argument that people engage in? Although it is by no means the only one, we ask this question with respect to a particular cognitive limitation having to do with multivariable causation, which we will argue to be particularly damaging in a context of discourse.

Inferring Causes

Reasoning about cause and effect is the most common form of reasoning humans engage in and the form most extensively studied by cognitive psychologists (Sloman, 2005). Criteria for inferring causes change during the first decades of life in ways that may seem paradoxical. Young children commonly regard an event as causal simply because it co-occurs with an outcome. They later adhere to more rigorous criteria and begin to distinguish causality from covariation and may even become able to eliminate potential causes via controlled comparison.

Surprisingly, however, young teens who have mastered controlled comparison are likely to attribute an outcome to a single factor, even when they have themselves just demonstrated that other factors present also affect the outcome (Kuhn, 2007, 2012). Moreover, the single factor to which they attribute causal power shifts across instances examined, whether or not prior

beliefs influence these attributions. In everyday reasoning unconstrained by consideration of specific evidence, a single favored cause is likely to suffice to explain a phenomenon. Overeating, for example, adequately accounts for obesity.

Further indication of a preference for single-cause explanations comes from a study by Gopnik et al. (2017). These authors studied causal inference patterns from age 4 through adulthood, reporting that 90% of 4-year-olds implicate an object merely present as causal in making a machine light up. By ages 12–14 and into adulthood this percentage dropped to below 40%, even when participants had witnessed cases in which two objects had been required jointly to produce the effect (and to less than 10% when they had not witnessed such cases). The remaining majority named only a single causal object. Gopnik et al. interpret this age difference as reflective of greater cognitive flexibility early in life. Yet given the evidence noted regarding a tendency to attribute an outcome to a single factor among older children, Gopnik et al.' s data may reflect simply the weak criterion of co-occurrence to warrant a causal inference early in life that with age becomes more demanding.

How Many Causes Produce an Effect?

The preceding evidence suggests that by adulthood people have a strong preference to explain an event as the result of a single cause. Since most real-world phenomena of interest are contributed to by multiple causes, such a tendency merits our attention. Our objective in what follows is, first, to ask how broadly this tendency extends to adults' thinking about everyday phenomena and, second, to consider its implications for discourse about such phenomena. Toward this end, we describe two impromptu studies and one more extensive, published study, concluding with a discussion of potential remedies.

The first impromptu study consisted of our asking a cross section of people at several Dunkin' Donuts coffee shops across a large city what had caused a specific event. We found only one previous study that had done anything similar. Strickland et al. (2017) asked participants from Mechanical Turk to "list as many causes as possible" for events such as "A woman is surprised." Respondents named a mean of 3.65 possible causes (vs. 2.50 for a physical event, e.g., "A window breaks"). We can't be sure, however, that these were

contributory versus alternative possible causes.[2] We therefore asked about a specific past event: Why did Jane Doe—a middle-aged, married, working woman from the Midwest—vote for Donald Trump in 2016? We encouraged respondents to elaborate the causes they identified, asking "What went into her decision?" and then "Is there anything else you can add?" To remove the possibility that respondents were constrained in their ability to envision possible causes, we showed them a list of possible causes but emphasized that these were simply for illustration and need not affect their response. Respondents were given $5.00 for their time.

Of 24 respondents, 17 named just one cause, five named two causes, and two named three causes.[3] Results differed little when we additionally asked why they themselves had voted for the candidate they did (or would have voted for, if they hadn't voted). Of the 24, 21 named a single cause, despite being prompted for anything else they might add.

Causal Reasoning and Discourse

Single causes are not sufficient to account fully for most real-world events. Kuhn and Modrek (2018) investigated possible implications of single-cause thinking for discourse. If a single cause is regarded as sufficient to explain an outcome, we hypothesized, alternative causes may be seen as contradictory, with implications for discourse: Either my cause or your cause must be the correct one. To assess this possibility, we constructed the simple three-item assessment in Box 14.1 and administered it to a cross section of 41 community adults.

In the three parallel items, option A (see Box 14.1) makes another causal claim, failing to address the initial claim and as a result not serving to address the stated objective of showing this claim to be wrong. Option B cites evidence with respect to an alternative sufficient cause; that is, the outcome may appear in the absence of the alleged cause due to another cause sufficient to

[2] In another experiment, Strickland et al. (2017) asked participants to choose a linear versus a converging diagram as representing such causes, ones unnamed. About two-thirds chose the converging diagram for human events versus one-third for physical events. Hence, in the abstract, respondents appeared to have some appreciation of contributing causes, more so in the human than in the physical domain.

[3] Of the seven naming more than one cause, however, four named as an additional cause that Jane's husband told her who to vote for, a cause external to Jane's own volition. The causes named for Jane's choice were diverse, most frequently "the country needed a change" ($n = 5$) and "to make America great again" ($n = 4$).

Box 14.1 Assessment Items

1. Some health officials have found cancer rates higher in cities than in outer areas. Dr. J. Rawls claimed tanning salons are to blame. Circle ONE piece of evidence that would be best to use if you wanted to argue he was wrong.
   A. Air pollution is a more likely cause of cancer in the city.
   B. Many people who don't go to tanning salons also get cancer.
   C. Many people outside the city also go to tanning salons and don't get cancer.
2. People from some countries have longer average life expectancy than people in others. Dr. F. Cole claimed a diet high in fish causes long life. Circle ONE piece of evidence that would be best to use if you wanted to argue he was wrong.
   A. Exercise is a more important cause of long life.
   B. People who don't eat fish often live to an old age.
   C. People who eat a lot of fish often live only to an average age.
3. Venezuela is a country with money trouble, unable to pay its bills. Dr. P. Garet claimed the cause was too many social programs to help people. Circle ONE piece of evidence that would be best to use if you wanted to argue he was wrong.
   A. Poor money management is a more likely cause of a country's money trouble.
   B. Some countries like Haiti have very few programs to help their people and Haiti has serious money shortages.
   C. Some countries like Sweden have many social programs and are not in money trouble.

Source: Kuhn and Modrek (2018).

produce it. Option B thus does not counter the claim that the initial factor is a cause. Option C, in contrast, does directly counter the claim that the initial factor is a cause since it cites evidence that this factor failed to produce the outcome.

Skilled reasoners might well regard such evidence as inconclusive in the absence of frequencies for all four cells (presence and absence of cause and presence and absence of outcome); however, untrained individuals rarely consider more than one or at most two of these cells (Schustack & Sternberg, 1981). None of our participants expressed such uncertainty; all read carefully, contemplated the three options, and chose one of them.

Of the 82% of respondents who showed a dominant response preference, about half chose option A most often, a quarter B, and a quarter C. Education level was the only factor associated with choice of the correct response, C. Replication with a comparable eight-item instrument yielded similar results, but the three-item version had nearly equivalent predictive power.

Do these individual differences in fact have the hypothesized implications for discourse? Box 14.2 contains the discussions of two pairs of individuals we recruited to participate in a dialog, one pair whose members both consistently preferred the correct but less frequently chosen option C on the three-item assessment (Box 14.1) and one whose members consistently preferred the more commonly chosen option A. All four had college degrees and had done some postgraduate work, thus reducing differences attributable to education level. The C-preference pair were participants in a graduate business course in strategic decision-making that involved a decision-making simulation. The A-preference pair were schoolteachers participating in graduate-level professional development training. Each pair was asked to choose from a list of suggested topics one on which they held opposing views and to engage in a dialog regarding it, trying to reach agreement if possible. The A pair chose whether the cause of teacher turnover is low pay or poor working conditions. The C pair chose whether educating people about the dangers of smoking or a high tax on cigarette purchases is most effective in reducing smoking.

The dialog of the C pair reveals several characteristics associated with high-quality discourse. First, both speakers cited actual or potential empirical evidence as the essential basis on which a claim is supported. Second, both understood that the two factors under discussion are not mutually exclusive alternatives—both may jointly and simultaneously contribute to the outcome ("I believe it's a combination of the two," P says explicitly)—and the dialog then turns to the relative efficacy of the two, again with an emphasis on empirical data as the basis for a judgment, recognizing that data may weaken as well as support a causal claim. Third, both P and N represent the dialog at a meta level—they make repeated reference to what they are doing and seeking to accomplish. When N acknowledges "You have a point," the subject is the dialog itself and the relation between the speakers' respective claims, rather than voicing of the claims themselves. P makes an even more ambitious meta-level effort to identify this relation: "I agree the government has a responsibility to stop people. I think we just disagree on the means by which they do this."

The dialog between A and O, by contrast, shows none of these characteristics. A and O alternate turns, each presenting their preferred causal candidates, by means of gradual elaboration, seeking to make their positions more convincing but without reference to evidence that would support the causal claim being advanced. Equally critical, neither directly addresses the

Box 14.2 A- and C-Pattern Sample Dialogs

Topic: Should smoking be reduced by educating people about its dangers or by charging a very high tax on purchase of cigarettes?

(C-Pattern Participants)

P: I favor education. Smoking is a personal decision. Something intrinsically very addictive and something people need to understand and make a decision for themselves. While I understand that people might vote, might purchase based off of their pocketbooks, you have to pay for smoking and if people really want something they're gonna find out how to do it probably to the detriment of other areas where they could be spending some of that disposable income.

N: I'm taking the other position that there should be a tax. There's plenty of evidence to suggest that smoking kills and the government has a responsibility to stop people hurting themselves.

P: I agree the government has a responsibility to stop people. I think we just disagree on the means by which they do this. And I'm going to point to two data points that I think rebut and actually state that raising taxes and making people decide based off of their pocketbooks has not been effective. I think the first thing we can talk to are a number of illegal drugs right now that are on the street. You see people who have very little money don't purchase food but they find the means to buy those drugs by any way possible. By the fact that there is a high price they're not only going to be purchasing them, to their detriment they're not going to be purchasing the things they need. That's my first argument.

N: Let me disagree with that. You have a point that people do buy illegal drugs. But on the other hand the government has a responsibility, and there are many areas where governments do take action to help people. Drugs is certainly one. There are a lot of other products people cannot buy because the government thinks it's bad either for them personally or for other people. And the fact that people are getting illegal drugs I think does not stop government's responsibility for trying to stop people from smoking by a high tax.

P: I don't think we disagree about whether it's the government's responsibility. It's the means by which they do it. I don't disagree it's the government's responsibility to educate, put programs in place. But I think the government should allocate those resources to education, not taxes.

N: I think people should be forced to pay. I think they should ban cigarettes altogether. But failing that, by making it really expensive to people is a good second best.

P: But if you had to pick one or the other, and the objective is to stop people from smoking, I believe it's a combination of the two. But if you had to pick one, is it higher taxes or education? And I think there's a lot of evidence . . . and I'm going to point to Denmark where I was watching a documentary where they actually legalized and kept the price the same—this was for some hard drugs—when they legalized it and they continued to educate the people—I don't have the data in front of me—but the amount of usage was reduced. This is one case study which might be contrary to the argument for raising taxes.

Box 14.2 *Continued*

N: People have been educated about the dangers of smoking for years. You even have to put on the cigarette box how dangerous it is to smoke. So it's pretty clear that doesn't happen. On the other hand, people do get worried about their pocketbooks and what they pay and I think that a higher price they have to pay will probably reduce their ability to smoke. There's probably been studies on that of when taxes have gone up in the past. I don't have that data in front of me but that would be something worth looking at.

P: I would tend to argue that between the 70s and 2016, if you were to look at the contributing factors, there's been a huge decrease in the rate of smoking in the last 30 to 40 years, as a per cent of population between the late 70s and 2016. If you were to try to dissect the factors that impacted that, you might find that in areas where there was a high tax, really there wasn't a decrease in smoking. So there's really no corollary* between a high tax and a decrease. But also schools that really focused on educating people, when in fact there was no increase in tax, you would find a decrease in smoking.

N: I'm sure there's data there and I think you're right, smoking has gone down over the years. But I think you have to look at the data and tease out of that data whether it was education or whether it was taxes. And I believe you will find that taxes had much greater effect than the level of education.

*[correlation]

Topic: Is the cause of teacher turnover low pay or poor working conditions?
(A-Pattern Participants)

A: So I think teachers are treated poorly for the amount of work they have to put in.

O: Maybe for some, but at the end of the day if salary was higher more teachers would probably stick around.

A: Not sure if I agree; it's how people treat you.

O: But you have to admit money incentivizes most people.

A: I think how you feel when you come to work and how appreciated you are is a stronger incentive.

O: So money has nothing to do with how happy or appreciated teachers feel?

A: I think working conditions, like administration and support, has a stronger impact on how we feel.

O: But salary would at least make more teachers stay.

A: Okay, teachers don't work for pay.

O: I didn't say that. I just think that higher salary would change the turnover rate.

A: Not sure if I agree; I mean, think of that lack of support from administrators.

O: Well there is need for more support from everyone.

A: Well yeah.

O: But turnover is high because many realize they aren't compensated enough for the amount of work they do.

A: Teachers do not get into this field because of wages.

Continued

Box 14.2 *Continued*

O: We're asked to do many other things besides just to instruct in the classroom and many are hardly making ends meet with the amount they get paid.

A: Okay, fine, but the reason for turnover is the way schools are run, not the money.

O: Salary change would make people want to stay.

A: Teachers go into the profession with a general idea of the salary but they can't predict the work conditions.

O: Not everyone knows what they're getting into.

Source: Kuhn and Modrek (2018).

other's claims, instead using one's conversational turn to elaborate one's own claim. Only at turn five does A first address O's claim of monetary cause by denying its causal status ("Teachers don't work for pay"), with O responding by reasserting its efficacy. This pattern occurs again, with A repeating the same denial ("Teachers do not get into this field because of wages"). Following another such repetition, A expresses the first counterargument to O's claim: "Teachers go into the profession with a general idea of the salary." Nor does either speaker evidence awareness that both their causes could be operating. Also absent is meta-level discourse about the exchange itself (beyond an unelaborated non-acceptance of the other's claim: "Not sure if I agree"). A and O may see no function of their dialog beyond one of airing their respective views, which they could have done outside a dialogic context. Their dialog thus reflects the failed or at least compromised interchange that may occur in the absence of the characteristics observed in P's and N's dialog.

The differences in the dialogs in Box 14.2, of course, do not establish definitively that the causal reasoning differences that distinguish the two pairs are the sole or even a major cause of the differing characteristics that appear in the quality of the dialogs. Other cognitive as well as personal–social differences between the individuals involved have not been eliminated. We did show that, overall, three- to four-member working groups in the business course whose members all showed a C pattern on the three-item assessment performed better in the business simulation that was the principal course activity than did those who showed lesser patterns. Still, further investigation of the differential characteristics of the dialogs is needed to make more definitive claims regarding their impact. In current work, we are examining the small-group, audio-recorded discourse of students in two different sections

of the business course, in order to evaluate the extent to which the groups of C-preference individuals are in fact more responsive to one another's statements and less shallow, relative to the groups lacking a preponderance of C-preference members, as appears to be the case.

Non-causal Thinking

In a second recent impromptu study (Kuhn & Cummings, 2018), we sought to examine whether similar differences would appear if the reasoning involved extended beyond causal reasoning to judgments involving principled decision-making and values. A total of 70 adults participated—41 community adults recruited from several urban public spaces and given $5.00 for their time and 29 students enrolled in a course at a nearby suburban community college. All indicated they were US citizens or legal residents. Each was asked, "What should be done about the problem of young people brought to the U.S. as children and now living in the U.S. illegally?" They indicated their position by pointing to a section on a line divided into seven segments. At one end appeared the phrase "Send Them Back" and at the other the phrase "Let Them Stay." They were then asked to explain the thinking underlying their judgments and finally to indicate how strongly they felt about the issue on a scale of 1–10.

Respondents were more likely to hold extreme views (– or + 2 or 3 on a 7-point scale from –3 to + 3)—0 of 69 (72%) did so (one did not choose a scale position). The remaining 28% indicated a moderate (–1, 0, +1) position. Respondents also felt strongly—a mean of 7.11 on the 10-point scale. Our main interest, however, was in how people justified their judgments. A judgment of value is, of course, a very different kind of judgment to justify than the causal ones considered previously. A fully adequate, comprehensive justification of one's position on this issue demands acknowledgment of at least two competing sets of considerations, namely, those of the society and its laws and those of an individual who did not knowingly violate them. We accordingly classified justifications into three categories: (a) those that included considerations on both of these sides, (b) those that identified multiple considerations but only on one side, and (c) those that justified their judgment by only a single consideration.

Most respondents, 51 of 70 (73%), fell into the single-justification category. (Student and community subgroups are combined as they responded

similarly.) Seven (10%) offered more than a single justification but only on one side, and 12 (17%) noted considerations on both sides of the issue.[4] Thus, the predominance of single-factor thinking parallels that found in the case of causal judgments.

Of further interest is the association of these reasoning types with both extremity of position and strength of feeling, where high affect conceivably might motivate thinking; also, however, it may constrain thinking, reducing attention to the possibility of alternative construals, whereas low affect in contrast may support more balanced thinking about an issue.

A contrasting direction of causality, from cognition to affect, is also possible. Under this interpretation, a simplistic single-factor representation regarding an issue allows strong affective endorsement because the cognitive representation includes no competing considerations. A more complex multi-factor representation, in contrast, tempers high affect due to awareness of additional considerations, especially if they weigh in opposing directions. On similar grounds, more complex, richer cognitive representations of an issue should be associated with less extreme positions on the issue.

Our findings were clear-cut. More complex thinking was associated with both less extreme positions and lesser affect. All 12 respondents who noted justifications on both sides indicated moderate (–1, 0, +1) positions on the issue. Of the larger group of single-justification respondents, the large majority, 90%, indicated extreme positions, with the multiple but one-sided group intermediate (but more likely extreme). Extreme positions were also associated with high affect (an average of 8.0) compared to moderate positions (an average of 4.79).

High affect could possibly enhance intellectual investment and energize thinking, leading to more nuanced, comprehensive thought; but there was no evidence that this was the case. Those reporting higher affective investment expressed less, not more, complex thinking. More likely, then, the evidence suggests, the potential influence of high affective investment is to constrain thinking.

[4] Factors mentioned most commonly on the "Let Them Stay" side were compassion, contribution to society, and consistency with US values. Factors mentioned most commonly on the "Send Them Back" side were legal, economic, and fairness (to legal immigrants).

Improving Reasoning as a Factor in Improving Discourse

A possible implication of the findings we have described here is that it may be productive to first seek ways to enrich people's cognitive representations of complex social issues, in contrast, for example, to seeking to temper their emotional investment. None of the studies that have been described were designed to weigh the relative importance of cognition versus affect in their effects on reasoning or discourse. Yet if thinking is sufficiently complex, it stands to constrain associated affect, for the reason that competing considerations are acknowledged and temper one another.

Related to this claim is the finding that simply asking people to explain the function of a common object (such as a toilet) tempers their assessments of this thinking, reducing their initial estimates of their understanding of the causal mechanisms involved (Rozenblit & Keil, 2002; Sloman & Fernbach, 2017). Yet a similar downward adjustment does not occur when people are asked for reasons for their sociopolitical views (Fernbach et al., 2013). Something more is required.

The more common cognition-focused approach to enhancing the quality of people's thinking on a complex issue has been to introduce factors that weigh on the opposing side, encouraging them to consider that perspective (Lao & Kuhn, 2002). To the extent that affective commitment is already high, however, this may meet with resistance and be difficult to do. An alternative we have begun to explore is to instead lead individuals to contemplate the implications and limits of the position they are already committed to.

In exploratory work with the immigrant topic, for example, we have explored introducing follow-up questions. If the respondent is favorable toward leniency, for instance, we ask whether the undocumented parents who brought their children to the United States should also be allowed to stay—the most frequent answer being yes. This leads to questions about others who might also be allowed to join them and ultimately about whether US borders need to be limited at all. A parallel set of questions posed to participants who are against leniency culminates in the question of whether the United States should prohibit immigration entirely. Although these efforts are at an exploratory stage, we are interested to see what the effects will be of engaging people in contemplating the full implications of their positions and how doing so may enrich their thinking and even temper their commitment to these positions.

Also relevant to the question of change is much developmental evidence that reasoning does improve with age and engagement (Crowell & Kuhn, 2014; Iordanou & Constantinou, 2015; Iordanou & Kuhn, 2020; Iordanou et al., 2016; Kuhn, 2018; Papathomas & Kuhn, 2017; Toplak et al., 201). As well as argumentative reasoning, this improvement encompasses causal reasoning both early in life (Walker & Gopnik, 2014) and at least into the second decade (Kuhn et al., 2015). Moshman (2018) in fact notes this fact as a limitation of Mercier and Sperber's (2011) account of argumentation as a tool to understood only in evolutionary terms. Conceptualizing reason simply as an evolved trait, Moshman asserts, draws attention away from its development, which involves both individual agency and social transaction.

Developmental potential, furthermore, exists specifically with respect to the cognitive achievement highlighted in this chapter, namely progression beyond a univariable model to recognition of multiple variables contributing to an outcome. Kuhn et al. (2015) found this progression achievable among young adolescents by engaging them over an extended period in inquiry activities involving causal investigation and inference with respect to phenomena that can be accounted for adequately only in multivariable terms. Whether a comparable result is achievable in the case of adults engaged individually or with one another in examining issues in which they may have considerable affective investment remains to be seen. The question, nonetheless, seems well worth investigating.

Finally, not to be neglected is the epistemological dimension (Iordanou, 2016; Iordanou et al., 2019). Individuals of any age engage in serious discourse only because they possess a set of intellectual values reflecting a commitment to the belief that such discourse is worth the substantial effort it entails (Kuhn, 2009; Kuhn et al., 2011). The roots of this epistemological dimension are to be found early in life, with the idea of beliefs as subject to revision rather than direct copies of reality (Iordanou, 2016).

Improving reasoning certainly will not by itself remedy all that is wrong with prevailing modes of discourse. The character of discourse, like the other phenomena that have been considered here, is influenced by multiple factors. Still, discourse can never be any better than the thinking that goes into it. Nor can it be any more productive. Of the two dialogs in Box 14.2, only the first likely served to enrich its participants' conceptualization of the issue discussed. Enhancing individual reasoning as a factor in improving discourse has not received as much attention as other factors, but it arguably warrants such attention by educators and more broadly. This is perhaps even

more the case at a time when people increasingly have become inclined not just to disagree with others' positions but to act on such differences.

References

Andersen, K. (2017, September). How America lost its mind. Atlantic Monthly.

Barbera, P., Jost, J., Nagler, J., Tucker, J., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*, 1531–1542.

Crowell, A., & Kuhn, D. (2014). Developing dialogic argumentation skills: A 3-year intervention study. *Journal of Cognition and Development*, *15*, 363–381.

Eliot, T. S. (2013). "The perfect critic." In T. S. Eliot (Ed.), *The sacred wood* (pp. 1–41). CreateSpace Independent Publishing Platform. (Original work published 1921)

Fernbach, P., Rogers, T., Fox, C., & Sloman, S. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*, 939–946.

Fisher, M., & Keil, F. (2014). The illusion of argument justification. *Journal of Experimental Psychology: General*, *143*, 425–433.

Gopnik, A., O'Grady, S., Lucas, C., Griffiths, T., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 7892–7899.

Iordanou, K. (2016). From theory of mind to epistemic cognition: A lifespan perspective. *Frontline Learning Research*, *4*, 106–119. https://doi.org/10.14786/flr.v4i5.252

Iordanou, K., & Constantinou, C. (2015). Supporting use of evidence in argumentation through practice in argumentation and reflection in the context of SOCRATES learning environment. *Science Education*, *99*, 282–311.

Iordanou, K., Kendeou, P., & Beker, K. (2016). Argumentative reasoning. In J. Greene, W. Sandoval, & I. Braten (Eds.), *Handbook of epistemic cognition*. Routledge.

Iordanou, K., Muis, K. R., & Kendeou, P. (2019). Epistemic perspective and online epistemic processing of evidence: Developmental and domain differences. *The Journal of Experimental Education*, *87*(4), 531–551.

Iordanou, K., & Kuhn, D. (2020). Contemplating the opposition: Does a personal touch matter? *Discourse Processes*, *57*(4), 343–359. https://doi.org/10.1080/0163853X.2019.1701918

Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education*, *91*, 710–726.

Kuhn, D. (2009). The importance of learning about knowing: Creating a foundation for development of intellectual values. *Perspectives on Child Development*, *3*, 112–117.

Kuhn, D. (2012). The development of causal reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 327–335.

Kuhn, D. (2018). *Building our best future: Thinking critically about ourselves and our world*. Wessex Learning.

Kuhn, D., Cummings, A., & Youmans, M. (2020). Is reasoning a fruitful path to changing minds? *Discourse Processes, 57,* 36–47.

Kuhn, D., & Modrek, A. (2018). Do reasoning limitations undermine discourse? *Thinking and Reasoning*, *24*, 97–116.

Kuhn, D., Ramsey, S., & Arvidsson, T. S. (2015). Developing multivariable thinkers. *Cognitive Development*, *35*, 92–110.

Kuhn, D., Wang, Y., & Li, H. (2011). Why argue: Developing understanding of the purposes and value of argumentive discourse. *Discourse Processes*, *48*, 26–49.

Lao, J., & Kuhn, D. (2002). Cognitive engagement and attitude development. *Cognitive Development*, *17*, 1203–1217.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57–111.

Moshman, D. (2018). Reasoning, logic, and development: Essay review of *The Enigma of Reason* by H. Mercier and D. Sperber. *Human Development*, *61*, 60–64.

Papathomas, L., & Kuhn, D. (2017). Learning to argue via apprenticeship. *Journal of Experimental Child Psychology*, *159*, 129–139.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521–562.

Schustack, M., & Sternberg, R. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, *110*, 101–120.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Sloman, S., & Fernbach, P. (2017). *The knowledge illusion: Why we never think alone*. Riverhead Books.

Strickland, B., Silver, I., & Keil, F. (2017). The texture of causal construals: Domain-specific biases shape causal inferences from discourse. *Memory and Cognition*, *45*, 442–455.

Toplak, M., West, R., & Stanovich, K. (2014). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental Psychology*, *50*, 1037–1048.

Walker, C., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science*, *25*, 161–169.

15

Knowing What Is Known

Emerging Insights into the Limits of Individual and Distributed Knowledge

Frank C. Keil and Kristi L. Lockhart

How do children come to understand what is known and what is knowable? This question differs from more classical epistemological questions about the nature of knowledge, including such issues as when knowledge requires certainty and whether it requires evidence. Here, we focus on how children envision the limits and boundaries of knowledge and how those views change with age. In particular, we are interested in what kinds of epistemological insights into knowledge boundaries emerge early in development, what ones take more time to manifest themselves, and why this developmental pattern unfolds as it does. In doing so, we will make the additional claim that this process is closely related to how children learn to master the division of cognitive labor that is intrinsic to all human cultures.

It might seem that young children should have very limited epistemological intuitions given the classic literature on metacognitive development. Thus, young children appear to have immense difficulties monitoring the contents of their own minds. They think they can recall far more items in memory in hidden picture tasks than they really can, apparently doing a dreadful job tracking what information they have recently acquired through experience (Yussen & Levy, 1975). They also sometimes report that they have always known something that they have just learned (Taylor et al., 1994). Yet, young children also seem to master several nuances of evaluating testimony offered by others, showing a sensitivity to confidence, consensus, and prior relevant domain competence, among other indicators (e.g., Harris, 2012). Thus, there seem to be areas where young schoolchildren have epistemological insights and others where they are much more challenged. Our central question focuses on the ways in which children are both impressively competent and surprisingly limited in terms of what they think is known and

Frank C. Keil and Kristi L. Lockhart, *Knowing What Is Known* In: *Reason, Bias, and Inquiry*. Edited by: Nathan Ballantyne and David Dunning, Oxford University Press.
DOI: 10.1093/oso/9780197636916.003.0016

knowable. We will argue that their epistemological strengths and weaknesses form a coherent account of how an understanding of knowledge develops.

We will address the question by considering three experimental approaches that explore different facets of knowing what is known. The first concerns judgments of what kinds of knowledge are acquired directly through first-hand experience and what ones are acquired indirectly from others. The second focuses on beliefs about learning potential or what we might call *knowledge futures*, namely how the course of knowledge acquisition is seen as unfolding in the future. The third focuses on appreciation of the virtue of ignorance, that is, how children come to appreciate that sometimes those who assert that something is intrinsically unknowable actually have greater insight than those who claim to have knowledge about the same topic. We will then see how, taken together, these developmental patterns are linked to the critical process of learning how to outsource knowledge effectively and reliably.

Direct Versus Indirect Knowledge Acquisition

One critical epistemological task is to learn what kinds of knowledge one could reasonably acquire on one's own and what kinds must involve input from other minds. To explore these intuitions, we developed a task that contrives a situation in which an individual can only learn about the world through direct experience (Lockhart et al., 2017). We used a "deserted island" scenario: An individual grows up alone on an island that has no other people or traces of human activity, a situation that is extremely benign such that the individual can be readily nourished and physically comfortable. This scenario may seem quite implausible to adults, but both child and adult participants found it easy to envision and reason about. In essence, we created a situation in which a person is a necessary autodidact (i.e., self-teacher), which ironically is the way much of early knowledge acquisition was traditionally discussed (Harris, 2001).

There are several reasons to believe that young children might find it especially challenging to make inferences about what kinds of knowledge the deserted island person could and could not acquire. It is, for example, well documented that young children have weaker source monitoring abilities, that is, abilities to keep track of the person or situation (e.g., a book or museum exhibit) that provided them with new information (Drummey &

Newcombe, 2002; Gopnik & Graf, 1988). To the extent that those abilities are limited, young children might not keep track of how they learned information and whether it was acquired from others or through their own direct experience.

It is possible that one facet of the development of source monitoring skills could involve children learning about the distinctive properties of direct versus indirectly acquired knowledge. They might need to build up a large set of instances of knowledge acquisition that are tagged as to whether the knowledge was acquired by the self or through others. This accumulated set of instances might then enable them to notice certain commonalities among the instances of the two kinds of knowledge and how they contrast with each other. For example, they might notice that directly acquired knowledge cases tend to involve directly perceivable entities that were encountered through interactions with the world. Although this is not universally true as one can learn something through an internal insight, such entities are certainly more central to directly acquired knowledge.

Alternatively, children improve in their ability to engage in a form of problem-solving in which one figures out in real time the epistemological logistics of acquiring a particular kind of piece of information. By *epistemological logistics* we mean the coordination of several components of a complex knowledge acquisition scenario, such as timing, point of view, working memory load, and information quality. Thus, one might realize that something typically happens too fast for a person to notice without the assistance of other observers (e.g., where different observers might focus on different parts of an event, such as a magic trick, and not have to incur costs of switching attention) or of technology (e.g., where a video recording can later be slowed down so as to be more easily interpreted by a single observer). In such cases, younger children might be expected to have a much weaker sense of these logistical challenges because they know less about perception, memory, and how information is structured in the environment. By this account, a major pattern of developmental change concerns learning about such logistics. Moreover, it contains many subcomponents such as a sense of historical/generational knowledge, of instrument-mediated knowledge (e.g., how some kinds of knowledge, such as that of germs, require specialized artifacts that can extend the ranges of perception and/or cognition), of the limits of human perception and memory, as well as of ways the complexity of the world might overwhelm human cognitive capacities. Because of the particular ways these challenges are manifested in the minds of children, it

seemed plausible that younger children would be prone to show what has been called an *individualism bias* (Gelfert, 2011), namely a tendency to assume that more information can be acquired on one's own than is really the case.

In order to determine whether children have different epistemic stances such as the individualism bias, in one set of studies we created different kinds of potential knowledge contrasts to explore children's intuitions about what the deserted island individual could know. The first set of contrasts involved *indirect versus directly* acquired knowledge, the second focused on *known versus unknown* forms of indirect knowledge as compared to direct knowledge, and the third contrasted direct knowledge that was relatively difficult to acquire versus effortless to acquire. Through these different kinds of contrasts we were able to use the deserted island scenario to probe how epistemological views change with increasing age.

Consider the first contrast, between direct and indirectly acquired knowledge. We asked children ranging in age from 5 to 10 years, as well as adults, whether or not the deserted island person would know such things as that "the rain comes from clouds in the sky" (direct), as opposed to whether the person would know that "the earth is round" (indirect). The direct items all involved either events that were directly observable or that were immediately apparent from internal mental experiences (e.g., "it's hard to think about two different things at once"). The indirect items were a varied set that ranged from distant historical information ("there used to be dinosaurs") to invisible entities such as those that are common in science and religion (e.g., "germs make people sick") to knowledge that requires instrumentation and inference ("stars are very hot"). In some cases, distant historical information and instrument-mediated information can overlap, as occurred when scientists realized that "hot" stars may not be hot at the present time but rather were hot when the light used to measure their temperature began its journey from those stars millions of light years ago. This overlap helps to highlight the contrast between learning to recognize some items as not directly learnable on the basis of intrinsic features (such as being historically remote or inaccessible to our senses) and others on the basis of how humans come to know about them (such as through certain instruments). This contrast illustrates more clearly how items can be indirect because of either of the two bases or because of a combination of both.

When children were asked to judge how likely it was that the deserted island person knew each of these things, we were surprised to find that even

the youngest children, kindergartners, thought that the person was much less likely to know the indirect knowledge. However, there was also a pattern of developmental change. For the direct items, all ages were virtually at ceiling levels in that they were highly confident that the person would know every direct item. In contrast, the youngest children, while judging indirect knowledge as less likely to be known than direct knowledge, nonetheless sometimes said that the person possibly knew the indirect knowledge items as well. This developmental trend revealed a bias toward greater knowledge attribution in younger children, a finding we have explored further in several later studies to be described in this chapter.

In short, young children are sensitive to the distinction between knowledge acquired from first-hand experience and knowledge that requires testimony or access to the products of other minds. But the finding that they sometimes attribute indirect knowledge to the deserted island person opens the possibility that their epistemological insights might be influenced by other factors not found in older children and adults. One such factor could be familiarity. Perhaps younger children consider how familiar the knowledge is in terms of their daily experiences. Many cases of indirect knowledge might be less frequently encountered in their normal lives than direct knowledge. To explore this possibility, we pitted three different kinds of knowledge against each other. For example, we asked children if the deserted island person knew "how to tell if it's raining" (direct), "how to ride a bike" (indirect-known), and "how to fly a helicopter" (indirect-unknown). We included both procedural knowledge (knowledge of how to do something, as illustrated here) and factual knowledge to see if the different kinds of knowledge mattered.

Once again, the youngest children were surprisingly precocious in their abilities. They judged the indirect knowledge items of both types as much less likely to be known than the direct ones. Moreover, there was no difference in judgments for the two kinds of indirect items. Familiarity didn't seem to sway them at all. In addition, the children once again showed the developmental trend in which the younger ages were more prone to say that the indirect items were known.

A different dimension of knowledge that would seem to have developmental consequences involves the challenges faced in acquiring some kinds of direct knowledge. That is, even if knowledge is technically accessible directly, in practical terms it still might be very difficult to acquire without external assistance from either other people or devices. Insights here would

seem to require knowledge of the pragmatics of situations that afford easy access to knowledge versus difficult access. Relatedly, an understanding of cognitive and computational limitations may be needed to sense that some kinds of direct experiences are logistically extremely difficult to notice and/or remember. For example, while directly available, it is a considerable challenge to keep track of all of one's eye blinks over the course of a week. To assess this contrast, we created direct easy-to-acquire knowledge items (e.g. "fish can't live outside of water") and direct difficult-to-acquire knowledge items (e.g., "boy birds have more colorful feathers than girl birds") and contrasted those with indirect impossible-to-acquire through direct experience items (e.g., "plants help make good air for animals to breathe").

With these contrasts, we began to see some developmental changes, in which older children and adults recognized that difficult-to-acquire items would be more likely to be unknown. However, this developmental trend was relatively modest in comparison to the larger contrast in which all ages judged the difficult direct items as less likely to be known than the easy direct items. Thus, even 5-year-olds were capable of considering the pragmatic and cognitive challenges of acquiring some kinds of information directly and using those estimated challenges to make inferences about the epistemological states of others. It was also clear that younger children were not quite as good at making such inferences and that we could surely contrive more intricate and subtler cognitive and pragmatic challenges that only older children, or possibly just some adults, could discern. These differences among different age groups might also be related to a greater reliance by younger children on reference to their own knowledge as a standard for evaluating knowledge in others (see also Birch & Bloom, 2007), but this developmental trajectory should not obscure the finding that even kindergartners can take into account pragmatic and cognitive challenges in reasoning about what others know.

Taken together, our deserted island studies show that, by the time children begin formal schooling, they employ a diverse set of skills to make inferences about what others feasibly know. Apparently, young children's weaker source monitoring skills and sparser understanding about human cognitive performance had little impact on these abilities. This raises the question why these abilities are so relatively well developed early on. Although the cause and effect relations are unclear, it may be related to the ubiquity of languages with evidentiality markers. More than 25% of the world's languages grammatically indicate whether a statement reflects one's first-hand experience

or information that was acquired second-hand (Fitneva & Matsui, 2009; Papafragou et al., 2007). This contrast and even more subtle ones are found in languages as diverse as Ersu, a Tibeto-Burman language, and Tariana, an Arawak language from northwest Amazonia (Aikhenvald, 2018). Thus, well before they start school, more than a quarter of the world's children must obligatorily indicate through their grammar whether or not they know something through first- or second-hand experience.

Ultimately, this precocious ability may reflect the importance of learning to navigate the division of cognitive labor. A sense of how others have acquired their knowledge might enable one to better understand "who knows what" in the world around them. Of course, many kinds of knowledge can be acquired either directly or indirectly, which might be thought to lead to an early bias to assume that if something could be learned directly, it was learned directly. Such a bias might be related to a tendency to overestimate not only one's present knowledge but also one's knowledge future.

Learning Potential and Knowledge Futures

Even if children are relatively well calibrated in terms of what is knowable through direct and indirect experience, they still might make systematic errors when estimating how much they currently know and how much they could know in the future. It is well documented that young children are much more optimistic about their futures in terms of traits and abilities (Boseovski, 2010; Lockhart et al., 2002, 2008). Thus, younger children think it much more likely that they, and their peers, will end up as adults with above-average levels in traits such as intelligence, attractiveness, and athletic ability. Could this also extend to their views about their future knowledge in terms of learning potential? There are at least two reasons to think it might not. First, even preschoolers are sensitive to the fact that adults can vary considerably in their degrees of expertise (Kushnir et al., 2013; Lutz & Keil, 2002). If children see this variability in knowledge as greater than that for physical and psychological traits, they might not think it is likely that most people's knowledge futures will be greater than average in all areas. Second, knowledge can be more transient than most physical traits. One can always forget things, learn new bits of information, and change one's mind. This greater instability of knowledge compared to traits might reduce optimism about what can be learned.

To explore these issues, we explored children's optimism about future knowledge in others' minds, in their own minds, and when the knowledge is strongly valenced (Lockhart et al., 2016). In addition, we asked whether optimism might vary as a function of the domain involved. It seemed plausible that even young children might consider some kinds of knowledge as less attainable in the future than others. This possibility is suggested by studies in which young children view some large domains of knowledge as intrinsically more difficult than others. For example, even young schoolchildren view psychological phenomena in general as easier to understand than phenomena in the physical and life sciences (Keil et al., 2010). Thus, children might have a more pessimistic view of future knowledge in the physical and life sciences.

We started our explorations with children's views of the knowledge futures of peers, assuming that their judgments of others might be less influenced by idiosyncratic personal life experiences. Children (5–7 and 8–10 years of age) and adults were asked how much others would know about phenomena relating to complex artifacts (e.g., "How much do you think John knows about all the inside parts that make up helicopters and how they work to make helicopters fly?"), biological processes (e.g., "How much do you think Tony knows about all the parts that make up trees and leaves and how they work to make leaves change color in the fall?"), non-living natural phenomena (e.g., "How much do you think Bill knows about all the parts that make up thunder and lightning storms and how these parts make thunder and lightning storms happen?"), psychological phenomena (e.g., "How much do you think Daniel knows about why some children are better liked than others and have more friends?"), and moral issues (e.g., "How much do you think Marty knows about when it's wrong to take other people's things without asking and why that's wrong?").

All the participants in this study were asked about how much a 5-year-old and that same person at age 35 would know. Asking about both ages provided a way of comparing what children think that young children would know with what they think adults would know. We were also curious about whether the youngest children would regard the same-aged peers (5-year-olds) much more positively because they were making judgments about their own age group.

Overall, we found a strong effect of age, with younger children being much more optimistic about what the 35-year-old would know. There was a secondary, smaller effect in which the youngest age group attributed more

knowledge to the 5-year-old than older children did; but the most important pattern was the much more optimistic outlook on what people would know as adults. There were also age and domain effects. In particular, adults thought that the 35-year-old would have the least knowledge about artifacts (entities made by humans) and somewhat less knowledge about natural kinds (naturally occurring classes of things), biology, and psychology. All ages thought the 35-year-old would have a great deal of knowledge about the moral issues. In contrast, the youngest age group thought that the 35-year-old would have a great deal of knowledge in all domains. The older children showed patterns relating to both age groups. They predicted more knowledge across all domains than adults, but they were also less optimistic and showed some differentiation across domains, rating artifacts as the hardest and moral topics as the easiest to know about.

How might judgments about one's own future knowledge differ from that about peers? Two possibilities arose. First, younger children, and perhaps all ages, might grant more knowledge to their future selves than to others, as part of the well-known Lake Wobegon effect (Dunning, 2011). Second, they might be very harsh about their earlier selves—thinking they have come a long way since that time—or they might think very highly of their earlier selves—thinking that they were always very knowledgeable. There are reasons to support both views. Young children do sometimes think they have always known something they just learned (Taylor et al., 1994), but their optimism might also lead them to think that they are tremendous learners who must surely know a great deal more than their earlier selves.

When children were asked the same questions about what they (as opposed to peers) knew as a 5-year-old and what they would know as a 35-year-old, the youthful optimism effect recurred. Thus, in contrast to the deserted island studies, here we again see how epistemological judgments about what is known can vary substantially with age. However, the results were also different from the cases where children evaluated others. All ages rated the self as having more knowledge than attributed to others, with the exception of a ceiling effect for ratings of the 35-year-old self by the youngest age group.

The youthful optimism effect might be manifested in two different ways. It could reflect a drive for a benevolent future in which only good things happen to an individual. Alternatively, one could have somewhat narrower beliefs about a hyper-competent future. In terms of future knowledge, this might mean either that the target adult only knew pleasant things or that the adult knew everything, both good and bad. Perhaps younger children are

more prone to believe that one would only want to learn about things that make one feel happy. A closely related issue concerns when in development one appreciates that having a piece of positive knowledge often also entails pieces of negative knowledge. For example, it seems that to have knowledge of what makes an agricultural crop succeed, one must also have knowledge of what makes the same crop fail.

To explore the influence of valence on epistemological judgments, we asked the same two age groups of children, as well as adults, what a 35-year-old would know about a topic in terms of both negative content and positive content. For example, one group of children might receive the following: "Sam is 35 years old and grown up. How much do you think he knows about why and how new lakes and rivers might form, creating a home where many animals and plants can grow and live?" (positive). In contrast, another group of children might receive the following: "Sam is 35 years old and grown up. How much do you think he knows about why and how lakes and rivers might dry up and disappear, leaving a desert where no plants or animals can live or survive?" (negative). Again, we gave examples of the five domains of artifacts, nonliving natural kinds, biology, psychology and morality.

The 8- to 10-year-old age group showed a modest valence effect in which they predicted greater positively valenced future knowledge. This was the first age at which spontaneous comments referred to pragmatic issues (e.g., "Why would someone want to learn something that made them feel sad?"). Younger children didn't appear to consider pragmatic factors and instead were near ceiling in optimistic forecasts for knowledge of both valences. Based on their comments, adults seemed to equate knowledge of both valences on grounds that one form of knowledge entailed the other, namely that one couldn't plausibly know the positive dimensions of a topic without also knowing the negative. The one exception to equal predictions about positive and negative knowledge by the younger children and adults was in the domain of morality, where all ages showed a tendency to think one would know more about the positive dimensions of a moral issue than the negative ones. Comments suggested that one wouldn't want to know negative moral things because one might be more tempted to engage in them. There might also have been an effect related to the deserted island studies in which participants of all ages were more likely to think that morality was directly acquired from first-hand experience, and they didn't want their protagonist to have had negative first-hand moral experiences.

The study on valence reveals multiple influences on judgments about one's knowledge future. We again see the early optimism. We also see the emergence of consideration of pragmatic issues at a later age. Finally, only in the adults does there seem to be a strong awareness of how negative and positive elements of knowledge might mutually entail each other.

Why should younger children be so much more optimistic about their knowledge futures? One argument concerns the adaptive value of being excessively optimistic early on (Bjorklund & Green, 1992; Lockhart et al., 2002). When children are young they learn at a very rapid rate and in many areas at the same time (e.g., simple rules of etiquette, how to find their way home, healthy foods to eat). They quickly progress from near total incompetence to serviceable knowledge or new physical skills. Greater optimism about what they can learn may help to motivate them in the face of otherwise seemingly overwhelming ignorance and incompetence. In addition, given that young children are usually protected from the consequences of their own ignorance by their caregivers, being overconfident as a preschooler about one's current and future knowledge is unlikely to cause much harm. It can be a very different story later on.

A second argument concerns the idea that it is quasi-rational for a young child to believe that one's future knowledge is virtually unlimited. It is very difficult to have a sense of future knowledge potential if one knows very little in the present; the less one knows, the harder it is to gauge one's ignorance (Dunning, 2011; Dunning et al., 2003). Given the tremendous early progress that children make in knowledge acquisition (e.g., learning new words) it might seem very reasonable to them that such a trajectory would continue for the foreseeable future. But again, one must also learn to outgrow this early optimism or risk real-life consequences that occur through extreme overconfidence.

Virtuous Ignorance

Even experts have gaps in their knowledge as well as beliefs about things they regard as intrinsically unknowable. Sometimes it is a virtue to be able to clearly state one's uncertainty or even that something is impossible to know. Indeed, in some instances a confident assertion of knowledge can actually be a clear indication of incompetence and lack of expertise. As adults, at least some of the time, we know enough to doubt certain kinds of knowledge

claims. For example, we would doubt a person who claimed to know exactly how many grains of sand there were in the entire world. We would rule out the possibility of that kind of knowledge for several reasons. We would see it as impossible to count all grains of sand everywhere in a feasible amount of time. We might realize that the precise number is intrinsically unstable as new grains are constantly created and others are destroyed. We might think that there is a built-in vagueness in the term "grain" that would cause indeterminacies in trying to decide whether something was a grain, a pebble, or a particle and that such decisions might even depend on contextual factors. But such an awareness would clearly depend on knowing a great deal about the world and how humans are able to gather information about entities in their environment.

In the deserted island studies, we saw that even young children were able to take into account factors indicating whether knowledge was acquired first- or second-hand; but the ability to doubt expert claims depends on much richer senses of plausible links between knowledge and the causal structure of the world. In addition, one must have a sense of randomness, scale, and uncertainty in various domains.

Finally, one has to reject other factors that would normally be indicative of possessing knowledge. For example, statements made with confidence are normally expected to be more likely true than those made with reservations. Given that young children are quite sensitive to confidence as a sign of possessing true knowledge (e.g., Jaswal & Malone, 2007), they might well find it difficult to suppress that indication because of other factors that argue against it, especially since young children have more executive processing difficulties in tasks where they must override a compelling piece of information because of other evidence (Jaswal et al., 2014). In addition, prior studies have shown that young children are less cynical about confident claims even when such cynicism is strongly suggested by the situation (e.g., Mills & Keil, 2005). For these reasons, we suspected that studies in which the ignorant person confidently claimed to have knowledge and the knowledgeable person claimed ignorance would pose a special challenge for younger children. We explored this possibility in two ways: claims of unjustifiable numerical certainty and claims of impossible predictive certainty. We contrasted these with claims of plausible numerical certainty and claims of feasible predictive certainty.

As the grains of sand example illustrates, some claims of numerical knowledge reveal incompetence because of multiple factors suggesting that

numerical precision is impossible. Other claims of precise numerical knowledge, however, seem to be well justified. As argued earlier, the ability to draw such distinctions might require considerable real-world knowledge. We explored how such an ability develops in a study with children in Grades K through 4 as well as with adults. We decided to look at more fine-grained age distinctions in this study because of our predictions of substantial change during the school years.

We presented children and adults with two experts and asked them to judge who was the "better expert." Previous to this choice, all children were given a brief example of what it means to be an expert (e.g., "So, someone could be an 'expert' in x if they really understand x and how it works really, really well"—examples of x were given). The task involved a series of questions posed to the "experts," followed by their answers. The following is an example of knowable numerical certainty:

> "If you count the number of windows in the White House, how many will you get?" "Expert" 1: There are exactly 147 windows in the White House. "Expert" 2: I don't know because it is not possible to answer that question precisely.

A case of unknowable numerical certainty is as follows:

> "If you count all the leaves on all trees in the entire world, how many will you get?" "Expert" 1: There are exactly 809,343,573,353,235 leaves on all trees in the world. "Expert" 2: I don't know because it is not possible to answer that question precisely.

Across a wide range of such examples, we observed a major developmental shift. For the knowable items, all age groups were at near ceiling levels in choosing the person who gave the precise answer as the better expert. In contrast, for the unknowable items, there was a marked developmental change, with the younger children picking the person who gave the precise answer and the oldest children and adults strongly preferring the person who said they didn't know.

We wondered whether the younger children might simply be having difficulty thinking about large numbers and perhaps did not realize how large those numbers were. To address that concern, in a second study, we looked at certainty about future events, some of which would likely be knowable and some of which were almost surely unknowable.

Knowable: "What colors will a rainbow have on April 4, 2721?" "Expert" 1: A rainbow will definitely have the colors red, orange, yellow, green, blue, indigo, and violet on April 4, 2721. "Expert" 2: I don't know because it is not possible to answer that question precisely.

Unknowable: "How long will the president's spouse's hair be, in inches, on February 17, 2033?" "Expert" 1: The president's spouse's hair will definitely be 15 inches long on February 17, 2033. "Expert" 2: I don't know because it is not possible to answer that question precisely.

In these cases, there were either no numbers involved or small, easy-to-grasp ones. The results almost perfectly mirrored those found in the first study on numerical certainty. All ages were near ceiling in their endorsements of the expert who made a specific knowable prediction. Again, in contrast, for the unknowable predictions, younger children strongly preferred the expert who made a specific prediction, while the oldest children and adults equally strongly preferred the expert who said they didn't know. In retrospect, this development shift should not be surprising. An understanding of what makes a prediction knowable versus unknowable requires a grasp of what are relatively stable versus transient causal patterns in the world. It would be surprising if young children were as sophisticated as adults in making such contrasts.

The development shifts found in numerical and predictive certainty studies were not just a result of weaker knowledge about world. In a follow-up study, we found that even when young children were able to judge one kind of knowledge as much more knowable than the other, they still often judged the confident expert for the unknowable items as better. It seems that they had great difficulty integrating information about confidence and knowability into a single coherent representation and that when the two were in conflict, confidence dominated. This dominance may be partly because their knowability intuitions were weaker as they depended on complex world knowledge that they had yet to acquire. Future studies are needed to understand the primary factors driving the developmental change.

Making Sense of the Overall Developmental Pattern

We have considered judgments of directly and indirectly acquired knowledge, beliefs about "knowledge futures," and appreciations of the virtues of

ignorance. Looking across the three sets of studies, we see cases of both early precocity and developmental lags in the ability to make mature judgments about what is feasible for others to know. In terms of detecting whether information was acquired either first-hand through direct experience or second-hand through testimony or supporting artifacts, even 5-year-olds were surprisingly sophisticated. In terms of optimism about future knowledge states, the second set of studies showed that younger children were much more optimistic. Indeed, they were so optimistic that there were few domain effects in the younger ages. At the same time, children showed some resemblance to adults in terms of self-enhancement effects, immunity from valence effects, and treating moral knowledge differently. Finally, in the studies pitting certainty against knowability, children showed the most dramatic shifts of all.

How can we account for these early versus late abilities? Several factors may be involved. First, there appeared to be powerful domain-general biases that apply to cognition far beyond epistemic judgments. These would include the early optimism bias and executive processing limitations on the ability to integrate information. As noted, the optimism bias has been found for beliefs about physical traits such as height, performance traits such as athleticism, and psychological traits such as shyness. In addition, the early optimism bias robustly occurs across diverse cultures (e.g., Lockhart et al., 2008). Apparently, the motivational value of seeing a very positive future is so strong that it trumps other factors.

Executive processing limitations are evident in many different areas of child development. Difficulties in integrating competing sets of information can be found in mathematical reasoning (Blair & Razza, 2007), theory of mind tasks (Sabbagh et al., 2006), and reading (Sesma et al., 2009), among others. Such challenges were especially evident in the virtuous ignorance tasks where younger children found it difficult to override confident but impossible-to-know assertions.

A different way of understanding what develops considers the adaptive value of the various biases that were studied. We argued earlier that the optimism bias may be highly functional because it encourages children not to give up in the face of their immediate failures when long-term gains seem so promising. In addition, youthful over-optimism may have far fewer side effects in young children because of a more protective caregiver environment. Indeed, this may be related to the tendency of younger children to see

boasting behavior as much more benign and for adults to be more tolerant of boasting in young children (Lockhart et al., 2018).

Adaptive or functional value may also help explain why, when confidence was pitted against plausible knowability, confidence won out in younger children. In general, adults and other seeming experts may be more likely to avoid deception and exaggeration when speaking to young children. This may be especially true because most of the adults that young children encounter are well-meaning individuals who have a deep interest in the child's welfare. In such situations, a confident assertion may be much more likely to be true than when a child is older and more likely to have discussions with peers and adult strangers who have less interest vested in the child having a positive outcome.

It is perhaps surprising that young children are able to figure out what people know despite knowing so little about how the world actually works. It is certainly possible to contrive cases where nuances of cognition and the causal structure of the world are elusive to younger children; but, across all our studies, the most reliable result was that this did not seem to hinder them very much. Young children do quite well with highly fragmentary and incomplete understandings of perception, cognition, attention, and how they all connect to real-world patterns. Their proficiency may suggest that all of us make such judgments quickly and without much cognitive load. This would certainly be consistent with the ease with which adults and children are able to use appropriate markers in evidentiary languages.

More generally, children's adeptness at making appropriate epistemological inferences with highly incomplete and fragmentary knowledge may presage the surprising extent to which adults make inferences in a similar manner. The "individualism bias" applies to adults as well, albeit to a lesser extent. Illusions of explanatory depth (Rozenblit & Keil, 2002) help to obscure the massive extent to which our abilities to reason about the world are dependent on what others know or external records of information created by humans. Instead of viewing young children as having disabling epistemological deficits, they may be better understood as revealing a framework of foundational abilities that enable all of us at any age to make knowledge inferences rapidly based on highly incomplete information.

Early emerging abilities supporting assessments of what others know may be at the heart of what makes us human, namely the ability to create and navigate the divisions of cognitive labor that occur in all cultures, and indeed define largely what it means to have a culture. Given the shallow and

incomplete nature of our knowledge, especially of the causal-explanatory kind, it is critical that we have a sense of what others know. Ideally, that sense should include intuitions about what kinds of things different individuals know as a result of their life experiences, the extent of knowledge in other minds, and the degree to which others approach maximum plausible knowledge in an area. Even 5-year-olds are able to partition up communities into different groups of experts who have distinctive mastery in understanding causal patterns that undergird such broad domains as physical mechanics or biology (Keil et al., 2008). Here, we have focused not on intuitions about such domains per se but rather on inferences about what is knowable given real-world contingencies. Such inferences are an essential part of mastering the division of cognitive labor. To defer to others appropriately, one needs to know whether one's own knowledge is adequate for one's needs even when the information is in an expert's domain and, if not, whether the expert is likely to have more of the relevant knowledge.

Conclusions

By the time they begin formal schooling, children have a diverse set of intuitions about what is likely to be known by themselves and others and what is likely to be knowable at all. Younger children do differ from adults by being more optimistic about how much they and others will know in the future, and they are much more swayed by high levels of unfounded confidence; but they nonetheless are able to assess what knowledge is acquired through direct experience and what must be acquired second-hand. In addition, through the early school years they soon master the abilities to override confidence and moderate their optimism, enabling them to benefit more richly from the divisions of cognitive labor that are such an essential part of all human cultures.

References

Aikhenvald, A. Y. (Ed.). (2018). *The Oxford handbook of evidentiality*. Oxford University Press.

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382–386.

Bjorklund, D. F., & Green, B. L. (1992). The adaptive nature of cognitive immaturity. *American Psychologist*, *47*(1), 46–54.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647–663.

Boseovski, J. J. (2010). Evidence for "rose-colored glasses": An examination of the positivity bias in young children's personality judgments. *Child Development Perspectives*, *4*(3), 212–218.

Drummey, A. B., & Newcombe, N. S. (2002). Developmental changes in source memory. *Developmental Science*, *5*(4), 48–62.

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. *Advances in Experimental Social Psychology*, *44*, 247–296.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87.

Fitneva, S. A., & Matsui, T. (Eds.). (2009). *Evidentiality: A window into language and cognitive development* (New Directions for Child and Adolescent Development 125). Jossey-Bass.

Gelfert, A. (2011). Expertise, argumentation, and the end of inquiry. *Argumentation*, *25*, 297–312.

Gopnik, A., & Graf, P. (1988). Knowing how you know: Young children's ability to identify and remember the sources of their beliefs. *Child Development*, *59*(5), 1366–1371.

Harris, P. L. (2001). Thinking about the unknown. *Trends in Cognitive Sciences*, *5*(11), 494–498.

Harris, P. L. (2012). *Trusting what you're told: How children learn from others*. Harvard University Press.

Jaswal, V. K., & Malone, L. S. (2007). Turning believers into skeptics: 3-year-olds' sensitivity to cues to speaker credibility. *Journal of Cognition and Development*, *8*(3), 263–283.

Jaswal, V. K., Pérez-Edgar, K., Kondrad, R. L., Palmquist, C. M., Cole, C. A., & Cole, C. E. (2014). Can't stop believing: Inhibitory control and resistance to misleading testimony. *Developmental Science*, *17*(6), 965–976.

Keil, F. C., Lockhart, K. L., & Schlegel, E. (2010). A bump on a bump? Emerging intuitions concerning the relative difficulty of the sciences. *Journal of Experimental Psychology: General*, *139*(1), 1–15.

Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive Science*, *32*(2), 259–300.

Kushnir, T., Vredenburgh, C., & Schneider, L. A. (2013). "Who can help me fix this toy?" The distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. *Developmental Psychology*, *49*(3), 446–453.

Lockhart, K. L., Chang, B., & Story, T. (2002). Young children's beliefs about the stability of traits: Protective optimism? *Child Development*, *73*(5), 1408–1430.

Lockhart, K. L., Goddu, M. K., & Keil, F. C. (2017). Overoptimism about future knowledge: Early arrogance? *The Journal of Positive Psychology*, *12*(1), 36–46.

Lockhart, K. L., Goddu, M. K., & Keil, F. C. (2018). When saying "I'm best" is benign: Developmental shifts in perceptions of boasting. *Developmental Psychology*, *54*(3), 521–535.

Lockhart, K. L., Goddu, M. K., Smith, E. D., & Keil, F. C. (2016). What could you really learn on your own? Understanding the epistemic limitations of knowledge acquisition. *Child Development*, *87*(2), 477–493.

Lockhart, K. L., Nakashima, N., Inagaki, K., & Keil, F. C. (2008). From ugly duckling to swan? Japanese and American beliefs about the stability and origins of traits. *Cognitive Development*, *23*(1), 155–179.

Lutz, D. J., & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child Development*, *73*(4), 1073–1084.

Mills, C. M., & Keil, F. C. (2005). The development of cynicism. *Psychological Science*, *16*(5), 385–390.

Papafragou, A., Li, P., Choi, Y., & Han, C.-h. (2007). Evidentiality in language and cognition. *Cognition*, *103*(2), 253–299.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind: A comparison of Chinese and US preschoolers. *Psychological Science*, *17*(1), 74–81.

Sesma, H. W., Mahone, E. M., Levine, T., Eason, S. H., & Cutting, L. E. (2009). The contribution of executive skills to reading comprehension. *Child Neuropsychology*, *15*(3), 232–246.

Taylor, M., Esbensen, B. M., & Bennett, R. T. (1994). Children's understanding of knowledge acquisition: The tendency for children to report that they have always known what they have just learned. *Child Development*, *65*(6), 1581–1604.

Yussen, S. R., & Levy, V. M., Jr. (1975). Developmental changes in predicting one's own span of short-term memory. *Journal of Experimental Child Psychology*, *19*(3), 502–508.

Index

For the benefit of digital users, indexed terms that span two pages (e.g., 52–53) may, on occasion, appear on only one of those pages.

Tables and figures are indicated by *t* and *f* following the page number

actor-observer differences, 37, 41–42, 46
adversarial systems of justice, 20
advice-taking, 306
advisor's paradox, 219
affect heuristic, 243–44
affective attitudes, 192–94
aggression and arrogance, 290
alcoholic rationality, 64
alternative facts, 56
ambiguous stimuli and visual prejudicial bias, 183–87
anger and arrogance, 290
Anton's disease, 139
applying knowledge, 210–11
arational attitudes, 192–94
argumentation theory, 305–6. *See also* cause and effect of arguments
arguments and information aggregation, 305–7
Aristotle, 309–10
arrogance, intellectual, 288–94
asking for advice, 219
associative judgments, 92
asymmetric overdetermination, 29–30
asymmetric self-perception, 44–45, 46
asymmetry evidence, 233–34, 238–39
attitude clarity, 284
attitude content, 281–82, 286
attitude extremity, 284
attitude function, 282–83, 286, 288–94
attitude in virtues and vices
  attitudes in social psychology, 276, 280–85
  cluster attitudes and, 285–88
  intellectual virtues and vices, 276–80, 288–94
  introduction to, 276–77
attitude object, 281, 282n.7
attitudes
  affective attitudes, 192–94
  arational attitudes, 192–94
  cluster attitudes, 285–88
  ego-defensive attitudes, 283
  emotional attitudes, 286
  emotive attitudes, 192–94
  explicit attitude, 284–85
  humility and, 287
  implicit attitude, 284–85
  instrumental attitudes, 283
  objective attitude, 256–57, 269, 272
  participant attitude, 256–57, 270–71
  propositional attitudes, 264
  social-adjustive attitudes, 283, 293–94
  value-expressive attitudes, 283
attitude strength, 284
attitude structure, 282, 286
Augustine of Hippo, 229
Austen, Jane, 113–14, 116–17, 120
austerity with rationality, 62
authority of rational agency, 266–67
autonomous judgments, 230
availability heuristic, 243–44

bad search challenge, 161–64, 167–68
Ballantyne, Nathan, 190–91
Bayesian models of visual prejudice, 180n.4
beginner's bubble of overconfidence, 217–18
behavioral disregard, 38–39
belief bias, 12, 30

beliefs
evidence-based beliefs, 255–56, 259–62, 264–65
explicit beliefs in information aggregation, 310–11, 312–17
false beliefs, 26–27, 28, 36, 88, 121–22, 215–16
irrational beliefs, 269
rational beliefs, 269
true beliefs, 14, 26–28, 70–71, 88, 265
believable rationalization, 114–18
benevolence and information aggregation, 304
Bethe, Hans, 245
bias. *See also* visual attention bias; visual prejudicial bias
belief bias, 12, 30
camera perspective bias, 133
cognitive bias, 29
confirmation bias, 90–91, 177, 180, 291–92
data bias, 12, 18
decision bias, 12
defined, 134
defined colloquially, 177
defined in psychology, 23–24
as dispositional, 21–25
formal bias, 177–79, 181
gender-biased judgment, 91
good bias, 92
group bias, 12
implicit bias, 20n.7, 50–51, 91, 93
imputations of, 44–45
inanimate bias, 12
individual bias, 12, 341–42, 354
information bias, 12
inquiry and, 30
intellectual virtues and vices with, 288–94
intrinsic bias, 28n.14
judgment bias, 12, 16–17, 21–22, 29
knowledge interaction with, 11, 29–32
liberal bias, 12
"My-side" bias, 79
opinion bias, 12
part-whole relations, 19–21
perception bias, 12, 55–57, 134–35
political bias, 19–20
prejudicial bias, 177–79, 183–87, 188–94, 197
procedure bias, 12, 13, 18
process bias, 12, 13, 15
race-biased judgment, 91, 181–87
racial bias, 183–87
reliability/reliabilism and, 11, 25–28, 178
sampling bias, 18
self-assessments of bias, 45
skill-based account of, 194–98
social bias, 12
status quo bias, 21
truth bias, 221–22
unbiased processes, 16–17, 20–23, 25, 26
verdict bias, 12
visual confirmation bias, 141–42
bias blind spot
breadth of, 47
causes of, 45–46
consequences of, 47–50
defined, 91
disagreement in, 48–50, 55–56
introduction to, 36, 44–45
knowledge claims and, 36, 44–51
reduction of, 50–51
biased assimilation, 243–44
blind spot in vision, 138–39
blind trust, 249
Boas, George, 1
body-worn cameras, 132–33
Brown, Michael, 129
bullshit characterization, 121–23

camera perspective bias, 133
Caplan, Bryan, 61
Carroll, Noël, 114–15
Cassandra quandary, 219–20
causal reasoning discourse, 327–33, 328*t*, 330*t*
cause and effect of arguments
causal reasoning discourse, 327–33, 328*t*, 330*t*
illusion of argument justification, 324
improved reasoning and, 335–37
inferring causes, 325–26
introduction to, 324–25

non-causal thinking, 333–34
number of causes to produce effect, 326–27
single-cause explanations, 326
censorship, 316
child knowledge acquisition. *See* knowledge acquisition by children
Churchland, Paul, 72
civil-law countries, 314–15
Climategate, 246–47
climate science research, 227–28, 239, 311–12
climate warming denialists, 246–47
closure principle, 95–96
cluster attitudes, 285–88
cognitive accessibility, 45
cognitive bias, 29
cognitive challenges of children, 344
cognitive empathy, 85, 85n.8, 86–87, 101–2, 103
cognitive mechanisms in information aggregation, 307–10
cognitive penetration, 150, 158–61
cognitive processes, 26, 28, 161, 169–70, 177, 192, 193, 205–6, 280–81
cognitive psychology, 35–36, 210, 325
Cohen, L. Jonathan, 77
collaborative learning, 315, 316–18
collective decision-making, 313–14
common-law countries, 314–15
communication and knowledge claims, 43
comparative advantage, 68–69
competence and information aggregation, 305
competence/performance distinction, 77
complex thinking, 334
computer mechanics, 78
Condorcet jury theorem, 302–3, 307–8
confidence and information aggregation, 304–5
confirmation bias, 90–91, 177, 180, 291–92
conflicting expert testimony, 228–29, 250–51
conflict-reduction strategies, 49
conformity-consistent evidence, 44
Confucius, 205, 206, 222
congeniality effect, 291–92
considered judgment, 30
"The Constellations Are Sexist" (McNeil), 74
Conway, Erik, 239
Correll task, 191–93, 195
cortically blind, 139
covert selection, 150–52, 165–69, 173
COVID-19 pandemic, 3
creative imagination, 113
crowd wisdom, 317–18
cues in information aggregation, 307–10
cultural cognition, 243–44

data bias, 12, 18
Davidson, Donald, 74
decision bias, 12
decision-making, 185–86, 313–14, 317–18
defensiveness and arrogance, 290
deference, defined, 229
deferential behavior, 293
deidealization, 64, 75–79
Delbanco, Andrew, 231
deliberative exclusivity, 109–12
deliberative weighing, 109, 112
demographic neutrality, 196, 197
Dennett, Daniel, 74
desirability of virtues and vices, 280
deviance assumption, 51–55
differential valuation, 38–39
diffusion analysis, 185–86
direct knowledge acquisition, 340–45
disagreement in bias blind spot, 48–50, 55–56
discriminatory associations, 89
dismissing expertise, 220–21
dispositional bias, 21–25
dogmatism. *See* perceptual dogmatism
Dunning-Kruger effect, 91, 93, 206–7, 209–10, 217–18, 219–20, 241–43

Eberhardt, Jennifer, 185
eclecticism, 71
educational memory, 63
efficient market hypothesis, 69
egalitarian values, 285–86
ego-defensive attitudes, 283
egoism, 65–66
eliminativism and rationality, 71–73
Eliot, T.S., 324

emotional attitudes, 286
emotional components of virtues and vices, 278
emotive attitudes, 192–94
epistemic authority, 266–67
epistemic equality, 86, 103
epistemic expert, 229
epistemic guilt, 96
epistemic norms, 165, 194, 195
epistemic trespassing, 220–21
epistemological judgments, 347, 348
epistemological logistics, 341–42
epistemology. *See also* perceptual dogmatism; Strawsonian epistemology
  of deliberation, 255
  folk epistemology, 49
  introduction to, 1–2, 6–7
  of promises and resolutions, 270–73
  reliabilism in, 26
  of self-knowledge, 263–70
  of testimony, 254–55, 257–63
  toxic epistemic environment, 246–48
errors of omission, 212–13
Evans, Gareth, 263–64
everyday paralogia, 216–17
evidence-based beliefs, 255–56, 259–62, 264–65
evidentialism, 178
experts/expertise
  assessment of, 5
  asymmetry evidence, 233–34, 238–39
  conflicting expert testimony, 228–29, 250–51
  deference, defined, 229
  epistemic expert, 229
  ignorant novices, 241–43
  introduction to, 227–32
  lessons for novices, 144
  naïve norms for evaluating, 144, 246
  partisan novices, 243–45
  reasonable deference, 229, 230, 232–41, 249–50
  toxic epistemic environment, 246–48
  in visual confirmation bias, 142
explicit attitude, 284–85
explicit beliefs in information aggregation, 310–11, 312–17
exposure control, 50
external factors of behavior, 41–42
extrospection, 36–44, 46, 53
eye-tracking, 139–40

facial recognition, 197
fake news, 56, 221–22, 247
fallacious arguments, 305–6
false answers, 216–17
false beliefs, 26–27, 28, 36, 88, 121–22, 215–16
false information gullibility, 221–22
Fodor, Jerry, 74
folk epistemology, 46, 49
forecast accuracy, 306
formal bias, 177–79, 181
formal legal truth, 143–44
formats in information aggregation, 310–12
Frankfurt, Harry, 121–23
free will, 39, 41–42, 254, 255
Fricker, Elizabeth, 257–58
Friedman, Milton, 75
functional value, 354
future behavior, 42, 43*f*, 44

Galileo, 76
Gardener, Dan, 317–18
Garner, Eric, 129–30
gender-biased judgment, 91
generational knowledge, 341–42
Goldman, Alvin, 238–39, 242–43
good arguments, 305–6
good bias, 92
good effects/outcomes from virtues and vices, 279–80
Greenspan, Patricia, 116
group bias, 12
group discussion, 307, 315–17
Guerrilla Girls, 142–43

Harris, Victor, 131
historical knowledge, 341–42
Hobbes, Thomas, 77
Hume, David, 93–94
humility and attitudes, 287
hypocognition, 213–15
hypoxia, 235–36

idealization, 64, 75–79
identity-defining group, 244
identity protection, 244
ignorance
  of ignorance, 207–10
  invisibility of, 210–18, 220–21
  personal ignorance, 218
  pluralistic ignorance, 36, 51–55, 54*f*
  rational ignorance, 218
  recognition of, 206–10
  self-implications of, 218–20
  societal implications of, 218–22
  virtuous ignorance, 349–52, 353
ignorant novices, 241–43
illusion of argument justification, 324
illusions of understanding, 208
imagination, 113, 114–18
imaginative pretense, 117
imaginative resistance, 107, 118–21
implicit attitude, 284–85
implicit bias, 20n.7, 50–51, 91, 93
imputations of bias, 44–45
inanimate bias, 12
inattentional blindness, 191
indirect knowledge acquisition, 340–45
individual bias, 12, 341–42, 354
inferring causes, 325–26
information aggregation
  arguments and, 305–7
  benevolence and, 304
  competence and, 305
  confidence and, 304–5
  crowd wisdom, 317–18
  cues and cognitive mechanisms in, 307–10
  defined, 301
  explicit beliefs in, 310–11, 312–17
  formats in, 310–12
  introduction to, 301
  juries in, 314–15
  majority opinion and, 303–4
  parliaments in, 313–14
  plausibility of, 303
  promise of, 302–3
  relevance of, 303–7
  schools in, 315
  science and, 315–17
inquiry
  attitudes and, 280, 282–83, 284, 285, 288, 293–94
  bias and, 30
  introduction to, 1, 2, 3, 6–8
  knowledge-conducive inquiry, 291
  norms of, 194n.14, 195
  by novices, 250
  rationalization and, 107, 112, 122–23
  by young adolescents, 336
institutional racism, 20n.7
instrumental attitudes, 283
instrumentalism, 64, 73–75
instrument-mediated knowledge, 341–42
intellectual arrogance, 5–6, 276, 280, 287, 288–94
intellectual confidence, 208
intellectual humility, 86–87, 100–3
intellectual servility/obsequiousness, 288–94
intellectual virtues and vices, 276–80, 288–94
internal cues in behavior, 37
internal factors of behavior, 41–42
International Survey of Painting and Sculpture exhibition, 142–43, 157n.10
interpersonal perception research, 134
intrinsic bias, 28n.14
introspection, 36–40, 37*t*, 46, 53
introspective weighting, 37–38
invisibility of ignorance, 210–18, 220–21
irrational beliefs, 159, 162, 269

judging others, 219–20
judgmental reliability, 238
judgment bias, 12, 16–17, 21–22, 29
juries in information aggregation, 314–15

Kahan, Dan, 243–45
Kahneman, Daniel, 87–88
knowledge
  asking for advice, 219
  assessment of, 5
  beginner's bubble of overconfidence, 217–18
  bias interaction with, 11, 29–32
  corralling the unknown, 214–15
  dismissing expertise, 220–21

knowledge (*cont.*)
Dunning-Kruger effect, 206–7, 209–10
everyday paralogia, 216–17
false beliefs, 215–16
gullibility of false information, 221–22
hypocognition, 213–14
ignorance of ignorance, 207–10
illusions of understanding, 208
introduction to, 6, 205–6
invisibility of ignorance, 210–18
judging others, 219–20
overclaiming knowledge, 207–8
recognizing one's own ignorance, 206–7
self-implications of ignorance, 218–20
societal implications of ignorance, 218–22
summary of, 222
unknowns and, 211
unknown unknowns, 212–13
knowledge acquisition by children
developmental patterns in, 352–55
direct *vs.* indirect acquisition, 340–45
introduction to, 339–40
learning potential and future knowledge, 345–49
summary of, 355
virtuous ignorance, 349–52
knowledge claims
bias blind spot, 36, 44–51
communication and, 43
free will and, 41–42
introduction to, 5–6, 35–36
introspection *vs.* extrospection, 36–40, 37*t*, 44
normalcy blind spot, 54–55
planning fallacy, 40–41
pluralistic ignorance, 51–55, 54*f*
self-knowledge *vs.* social-knowledge, 36–39, 38*t*
self-righteousness, 39–40
social influence perceptions, 44
strategy implications, 44
knowledge futures, 340, 345–49, 352–55
knowledge platitude, 31–32
Kornblith, Hilary, 96–97

Lake Wobegon effect, 347
language acquisition, 344–45
learning potential and future knowledge, 345–49
liberal bias, 12
liberal democracy, 230–31
linguistic illusion of inconsistency, 64
Lorenz, Konrad, 70–71
Lumiere, Auguste, 136–37

majority opinion and information aggregation, 303–4
Manhattan Project, 245
Marušić, Berislav, 255, 258–62, 270–73
mass shootings, 131–32
maximizing/minimizing inconsistency, 75–79
McNeil, Leila, 74–75
Mechanical Turk sample, 53–54
medical decision-making, 317–18
memory/memories
as decaying sense, 77
educational memory, 63
of information, 214
knowledge and, 339–40, 341–42
rationality and, 60, 70, 77, 78
of visual stimuli, 210
Mercier, Hugo, 79
meta-cognitive tasks, 205–6
Minsky, Marvin, 78
mitigating responses, 95, 98–100
Montaigne, Michel de, 210–11
mood swings as functional, 70
moral incoherence, 118–21
Moran, Richard, 255, 258, 261–62, 263, 264–67
Morris, Errol, 133
motivated reasoning, 244–45
motivational components of virtues and vices, 279
multi-track dispositions, 278
Mussolini, Benito, 133
"My-side" bias, 79
The Myth of the Rational Voter (Caplan), 61

naïve norms for evaluating expertise, 144, 246
naïve realism, 7, 39, 137–38, 141–42, 155–58

National Science Foundation, 221
natural selection, 61, 79
neuroscience of emotion, 115
Newtonian physics, 153–54
New York's Museum of Modern Art, 142–43
Nobel Prize, 245
non-causal thinking, 333–34
nonmonotonicity of imaginative blockage, 119–20
non-naïveté, 109, 110–12
normalcy blind spot, 54–55
normative standards, 23–24
norm of reasonableness, 83
North Atlantic Treaty Organization, 228
notion of interest, 23–24
nuclear war, 228

Obama, Barack, 215
objective attitude, 256–57, 269, 272
objective condition, 47–50
objective information availability, 186–87
objective reality, 55–57
O'Neal, Paul, 132–33
"one thought too many" argument, 262n.9
opinion bias, 12
Oreskes, Naomi, 239
other-knowledge, 39, 46
other-race face expertise, 190, 195
overclaiming knowledge, 207–8
overt cognitive selection, 150–52
own-race face expertise, 190, 195

Paddock, Stephen, 131–32
panpsychism, 71–72
parliaments in information aggregation, 313–14
participant attitude, 256–57, 270–71
part-whole relations, 19–21
past experiences matter challenge, 170–72
Pauling, Linus, 228, 240
Payne, Keith, 185
perception bias, 12, 55–57, 134–35
perceptual accuracy, 134, 155, 155n.7
perceptual dogmatism
  bad search challenge, 161–64, 167–68
  cognitive penetration and, 150, 158–61
  covert selection, 150–52, 165–69, 173
  covert selection and, 150–52, 165–67
  defined, 153–54
  direct challenge and, 164–69
  indirect challenge to, 169–73
  introduction to, 4–5
  naïve realism and, 155–58
  perceptual experience challenge, 4–5, 154, 155–58
  prima facie justification and, 153, 173
perceptual experience, 4–5, 154, 155–58
perceptual judgments, 156
personal ignorance, 218
personality traits, 286–87
perspective-taking, 43
planning fallacy, 40–41
Plato, 65, 229
plausibility of information aggregation, 303
plausible knowability, 354
pluralistic ignorance, 36, 51–55, 54*f*
political bias, 19–20
Politics (Aristotle), 309–10
post-hoc rationalization, 108
"post-truth" era, 55–57
pragmatic challenges of children, 344
pragmatic encroachment, 163–64, 167–69, 173, 178n.2
precocious ability, 345
prejudicial bias, 177–79, 183–87, 188–94, 197
previolation rationalizations, 108
prima facie justification, 153, 173
principle of charity, 62
priority of processes hypothesis, 11, 12–19, 21n.8
probability of unknowns, 211
procedure bias, 12, 13, 18
process bias, 12, 13, 15
propositional attitudes, 264
psychological egoism, 65–66
psychology
  bias, defined in, 23–24, 28
  cognitive psychology, 35–36, 210, 325
  introduction to, 1–2, 6–7
  of knowledge, 205–6
  social psychology, 35–36, 134, 276, 280–85

public discourse reasonability, 82–93, 94–95, 96n.25, 98–99, 101–2 103–4, 324

quarantining of pretense, 120–21

race-based facial expertise, 191–92, 195
race-biased judgment, 91, 181–87
racial bias, 183–87
Radford, Colin, 71
radical eliminativism, 72
rational beliefs, 269
rational ignorance, 218
rationality
  eliminating ingredients of, 71–73
  idealization *vs.* deidealization, 64, 75–79
  inconsistency about, 60–64
  inconsistency prognosis, 79–80
  maximizing as tautology, 66–69
  maximizing/minimizing inconsistency, 75–79
  maximizing without restraint, 65
  overview, 3–6
  as tool, 73–75
  yo-yo attributions of, 70–71
rationalization
  as believable, 114–18
  bullshit characterization, 121–23
  deliberative exclusivity, 109–12
  deliberative weighing, 109, 112
  of dishonesty, 120
  imaginative resistance and, 118–21
  model of, 109–14
  point of, 107–9
rationalizing, defined, 107
reasonable deference, 229, 230, 232–41, 249–50
reasonableness
  bad-bias argument, 93–104
  cognitive empathy, 85, 85n.8, 86–87
  defined, 82–83
  epistemic equality, 86
  intellectual humility, 86–87
  introduction to, 1–3, 4, 7–8, 82
  norm of reasonableness, 82–87
  as public discourse norm, 82–87
  reflection in, 84, 87–93
  response in, 84, 87–93
reasoning improvement and cause and effect, 335–37
recognizing one's own ignorance, 206–7
reflection in reasonableness, 84, 87–93
reflective equilibrium, 30
relevant base rate information, 40–41
reliability account, 236, 238
reliability condition, 240–41, 248–49
reliability/reliabilism
  bias and, 11, 25–28, 178
  covert selection and, 165–66, 165–66n.18
  in epistemology, 26
  judgmental reliability, 238
  of visual experience, 189
response in reasonableness, 84, 87–93
Roman Senate, 313–14

salient normative standard, 23
salient symmetry standard, 24–25
same-race face expertise, 189–90
sampling bias, 18
Samuelson, Paul, 62, 68, 69
schizophrenia, 135
schools in information aggregation, 315
science and information aggregation, 315–17
scientific method/standards, 221
Scott, Timothy, 131
selective exposure effect, 291
self-assessments of bias, 45
self-enhancement, 352–53
self-esteem, 290, 291–92, 293
self-implications of ignorance, 218–20
self-knowledge, 36–39, 38*t*, 263–70
self-other asymmetry, 38–39
self-perception, 38*t*
self-righteousness, 39–40, 116–17
sender/receiver conflicts, 304
Sense and Sensibility (Austen), 113–14, 116–17, 120
sensory imagination, 113
servility/obsequiousness, intellectual, 288–94
Siegel, Susanna, 188, 190
simple account, 234–35
single-cause explanations, 326
skepticism about reasonableness, 93–95

skill in visual prejudicial bias, 194–98
Smith, John Alexander, 231
social-adjustive attitudes, 283, 293–94
social bias, 12
social influence perceptions, 44
social intuitionist model of moral reasoning, 90
social perception, 37, 38*f*, 38*t*, 44–45
social psychology, 35–36, 134, 276, 280–85
societal implications of ignorance, 218–22
Socrates, 65, 77–78
Sperber, Dan, 79
"spill over" effect, 116
Stanley, Jason, 309
Stapp, John Paul, 227
status quo bias, 21
stereotyping, 89, 188–90, 192–93, 198
storing knowledge, 210–11
Strawsonian epistemology (Strawson, Peter)
  on freedom, 256–57
  introduction to, 254–56
  self-knowledge and, 263–70
  summary of, 274
  testimony and, 254–55, 257–63
subpersonal. relevant processing, 161
substantive truth, 143–44
Superforecasting: The Art and Science of Prediction (Tetlock, Gardner), 317–18
Surowiecki, J., 317–18
Swift, Jonathan, 227
synchronic inconsistency, 71

Tahitians of the Society Islands, 213
testimonial knowledge, 264
testimony, 228–29, 250–51, 254–55, 257–63
Tetlock, Philip, 317–18
"The Tower of Goldbach" (Gendler), 118–21
toxic epistemic environment, 246–48
trivial tautology, 68
true and false paradigm, 14
true beliefs, 14, 26–28, 70–71, 88, 265
Trump, Donald, 208
trumped incentive, 110–11
trust/trustworthiness, 249, 260–61, 311
truth bias, 221–22
truth-related norms, 244

unbiased processes, 16–17, 20–23, 25, 26
unknowns
  corralling of, 214–15
  knowing and, 211, 214–15, 352
  probability of, 211
  unruly unknowns, 211
  word game unknowns, 212
unknown unknowns, 212–13
unraveling responses, 95–98
unreasonable deference, defined, 229
unreliability and bias, 25–28
unrepresentative sample, 18
unruly unknowns, 211
unwarranted confidence visual attention, 136–41
US Chess Federation, 242

vaccine development, 3
valence effect, 348–49
value-expressive attitudes, 283
van Fraassen, Bas, 272
verdict bias, 12
video-editing technologies, 247
video evidence limitations, 131–33
virtues and vices. *See* attitude in virtues and vices
virtuous ignorance, 349–52, 353
visual attention bias
  error in visual experience, 133–36
  introduction to, 129–31
  limitations of video evidence, 131–33
  naïve realism, 7, 137–38, 141–42
  unwarranted confidence, 136–41
visual confirmation bias, 141–42
visual prejudicial bias
  ambiguous stimuli and, 183–87
  arational, emotive attitudes, 192–94
  candidate criteria, 188–94
  challenges with, 179–87
  face perception and, 181–83
  introduction to, 177–79
  irrational transformations, 188–89
  neglect of information and enquiry, 189–92
  skill in, 194–98
  stereotyping and, 188–90, 192–93, 198

vitamin C therapies, 228
voter irrationality, 61

well-being components of virtues and vices, 279
Williams, Bernard, 262n.9
Wilson, Darren, 129
The Wisdom of Crowds (Surowiecki), 317–18
word game unknowns, 212
World Economic Forum, 2

youthful optimism effect, 347–48
yo-yo attributions of rationality, 70–71