INTRODUCING IMPLICIT BIAS: WHY THIS BOOK MATTERS

Erin Beeghly & Alex Madva

Why do we do what we do? What makes humans tick? We often like to think that our actions are shaped by our choices, and our choices are mostly shaped by careful and well thought-out deliberations. However, research from the social and cognitive sciences suggests that a wide range of automatic habits and unintentional biases shape all aspects of social life. Imagine yourself walking through a grocery store. The smaller the floor tiles, the slower people tend to walk. The slower people walk, the more they buy. Hidden biases like this are well known to marketers and consumer psychologists. Yet store shoppers do not notice the ways in which the flooring affects their walking patterns or spending decisions (e.g. Van Den Bergh et al., 2016; see also Brownstein, 2018).

Examples such as this are only the tip of the iceberg. Increasingly, social scientists cite "implicit" or "automatic" mental processes to explain persistent social inequities and injustices in a broad range of contexts, including educational, corporate, medical, and informal "everyday" settings. Implicit biases have been invoked to explain heightened police violence against black US citizens, as well as subtle forms of discrimination in the criminal justice system and the underrepresentation of women and people of color in the workplace. Across the globe, employees are now required to attend implicit bias trainings, where they learn about these biases and their detrimental effects. A popular buzzword, "implicit bias" was even discussed by Hillary Rodham Clinton and Donald Trump during the 2016 U.S. Presidential debates (Johnson, 2016).

At the same time, critics of implicit bias research raise red flags. A key challenge—voiced by theorists of race, gender, disability, and economic inequality—maintains that the focus on individual psychology is at best irrelevant to, and at worst obscures, the more fundamental causes of injustice, which are institutional and structural (Ayala-López, 2018; Haslanger, 2015; Payne and Vuletich, 2017). Consider racial segregation in housing. To this day, neighborhoods in the United States are often racially segregated (for visualization of census data, see Cable, 2013). One could try to explain the phenomenon by citing individual preferences, choices, or beliefs, including racist beliefs, or a preference, perhaps common among people of color, to avoid discrimination from other racial groups (Feagin and Sikes, 1995). But such explanations may cover up the most important parts of the story. Throughout the twentieth century, the U.S. government created and maintained racial segregation through federal (as well as local and state) policy and laws. U.S. courts legally upheld policies that promoted residential segregation well into the 1970s, as did local police. The issue is this. If we center our explanation on individuals' beliefs and preferences regarding where to live, we have offered an explanation that hides the deepest causes of racial segregation and, in doing so, obscures our government's complicity in injustice (Anderson, 2010; Rothstein, 2018; Shelby, 2016).

This is not the only pressing criticism on the table. Some have argued that traditional psychology mistakenly characterizes biases as entirely "in the head" whereas biases may instead be actively embodied and socially constituted (Ngo, 2019). Other theorists have expressed doubts about the quality of the scientific research on implicit bias and the power of implicit bias to predict real-world behavior (Hermanson, 2017; Oswald et al., 2013; Singal, 2017).

Criticism comes from all sides. On one side of the political spectrum, implicit bias research is perceived as nothing more than a way for liberals to justify their political agenda by

claiming that it's backed by science—even though the underlying science is, these critics allege, flimsy (Mac Donald, 2016, 2017). On the opposite side of the political spectrum, implicit bias research is perceived as an evasion of the fact that old-fashioned, explicit bigotry never went away (Blanton and Ikizer, 2019; Haslanger, 2015; Lauer, 2019). Administrators only fund implicit bias trainings, this criticism goes, because people who run these trainings are paid to assuage participants' guilt and make them feel good. ("You're not racist. Even good people are implicitly biased!") Moreover, by paying lip service to the value of equality, they protect these institutions against lawsuits alleging discrimination and sidestep more impactful changes to their workplaces.

In this historical context, the present volume was conceptualized. As editors, our ambitions are big. Teaching and outreach are part of our mission. In our classrooms, we found ourselves wanting accessible resources, in particular, readings that could be used in contexts where not everyone had an in-depth background in social psychology, political theory, or philosophy of mind. Much of the existing literature on implicit bias is incredibly technical. Theorists tend to explore the implications of implicit bias for sophisticated, longstanding debates regarding, for example, moral responsibility or the cognitive architecture of the mind. Others try to persuade readers that implicit biases are important to specific debates within, for instance, the nature of perception. We ourselves have written technical papers like this! However valuable, these writings often miss the mark when it comes to introducing students to bias and its implications. We also found that more accessible discussions of implicit bias ended up sacrificing accuracy and rigor. What we wanted for our students was the best of both worlds: conceptually rigorous, empirically informed work written in a non-technical and accessible style, using engaging examples. We also wanted our students to have a big-picture sense of why bias matters, a goal that we thought would be best served by a collection that covered a broad swath of conceptual, ethical, and political terrain. The text we envisioned did not shy away from controversy. We needed our students to understand and reflect on the criticisms surrounding implicit bias, as well as how one might answer them by adopting a less simplistic understanding of implicit bias and its role in maintaining injustice.

With these goals in mind, we sought out contributors. All the authors in this volume are philosophers by training. They are also deeply immersed in the psychological and social-scientific literature on bias, and their chapters have been carefully peer-reviewed by other experts in their respective fields. Some specialize in the philosophy of mind and cognitive science. Some work in epistemology—the branch of philosophy that deals with knowledge. Some are ethicists. Others work in political and social theory. Each has their special area of expertise and methodological commitments, and together they represent a wide array of social backgrounds. Yet all of our authors share something in common. They write in a non-technical style, using relatable examples that help readers better understand what implicit bias is, its significance, and the controversies surrounding it.

Think of this volume as a guide to the territory. It is interdisciplinary through and through: bringing the philosophical perspective of the *humanities* together with the perspective of the *social sciences*. Each chapter focuses on a key aspect of bias. A glossary at the volume's end defines important terminology. Web resources are available, which provide food for thought and connections to wider cultural resources, including podcasts and movies. For those who want to dive deeper into the philosophical or empirical research, our authors have provided suggestions for additional reading at the end of each chapter. For teachers who want to use this

volume in educational settings, we also offer study questions. We want this book to be useful to you, no matter who you are and what you want out of it.

To that end, the book is organized into three parts. Part 1 explores what implicit bias is. It also examines a variety of critical arguments that suggest implicit bias does not really exist, or does not help to explain social injustice, or is based on shoddy scientific evidence. Part 2 focuses on questions surrounding knowledge and bias. One of the key aims is to document the myriad of ways in which biases impede knowledge. However, it also broaches the question of whether biases track truth and are reliable in some cases. Part 3 examines key practical, ethical, and political questions surrounding implicit biases. In this part of the book, our authors build on rich traditions of scientific research to ask questions that go beyond psychology, or that don't lend themselves to straightforward scientific investigation. For example, they ask whether—and how—we can be held responsible for our implicit biases, and how we ought to structure society to combat implicit biases.

Significantly, no one in this volume argues that implicit bias is somehow the most important target for social-justice efforts. However, what emerges throughout these chapters are the complex ways that implicit bias connects to other issues. Return to the example of racial segregation in the United States. While institutional tools—including federal law, the criminal justice system, and local police—were consistently deployed throughout the twentieth century to create and maintain residential segregation, one ought to ask why lawmakers, police, and many of the people whom they served supported white-only residential spaces in the first place. Who created these policies? How were they responsive to market pressures, if at all? Who pushed back against them? Individual psychology has to be invoked if these questions are to be answered fully (e.g. Alexander, 2012; Enos, 2017; Lopez, 2017; Payne et al., 2010; Pettigrew, 1998). One could never get the whole story about why segregation succeeded for so long—and persists even today—unless one talks about individuals' beliefs, desires, aversions, preferences, and so on. Segregation fed and was fed by biases, many of which were explicit prejudices but some of which were subtler, a fact to which Martin Luther King Jr.'s criticisms of white moderates attest (King Jr., 1963). Implicit bias is thus one piece of a complex puzzle.

Another way to highlight how implicit bias is one piece of a complex puzzle is to consider how it relates to explicit bias. One concern about implicit bias research is that it is out of touch with current social and political realities, which have witnessed a surge in intergroup political division and openly endorsed bigotry, including violent, white supremacist rallies and mass shootings in the United States as well as far-right, neo-Nazi parties gaining ground worldwide. Readers might understandably look at these trends and ask: why are we still talking about implicit bias? However, implicit bias research is surprisingly well-positioned to help us understand what's going on (Madva, 2019). First, consider findings from developmental psychology. Researchers find that infants and young children form social biases very early on, and even tend to openly report racial preferences through the age of six (Dunham et al., 2008). Ten-year-olds, however, become less likely to report explicit biases, while adults are much less likely still. Meanwhile, overall patterns of implicit bias tend to remain stable all the way from childhood to adulthood. These patterns suggest that children form explicit biases early on, but then gradually learn that these biases are wrong, and not OK to say out loud. Seen in this light, implicit biases are like a "residue" left over in people raised to endorse anti-prejudiced values even while they are immersed in a broadly prejudiced society (Payne et al., 2019).

In adulthood, however, this implicit residue of prejudice continues to affect behavior—and comes to function like a "sleeper cell" of bias, waiting to be activated when social norms

change. Studies find that when you plunk an implicitly biased person into a social context where authority figures and peers promote prejudiced norms and values, then *their implicit biases become explicit once again* (Crandall et al., 2002; Lee et al., 2017; White and Crandall, 2017). It takes very little, it turns out, for the implicit to bubble up into the explicit, and for suppressed prejudices to become openly endorsed and acted upon. In other words, explicit bias feeds and is fed by implicit bias (Gawronski and Bodenhausen, 2006; De Houwer, In press). One wouldn't appreciate complexities like this from the simplistic portrayals of implicit bias one finds in popular media, but the sophisticated studies being published at a rapid pace in top social-science journals have much to say about unfolding social and political events. When you dig deep into the nature of implicit and explicit biases, part of what emerges is their interconnections, and the fuzzy boundaries between them. In this vein, some of the chapters in this volume will consider bias in general, rather than focus solely on its more implicit forms.

What can you expect when you read this volume, more specifically? In the rest of this introduction, we'll provide a thumbnail sketch of each chapter. If you don't like spoilers, you can skip this section. However, if you'd like a more concrete sense of what our authors will be discussing in each chapter and the scope of concerns covered in this volume, keep reading.

Part 1

Knowing what to do requires knowing what we are up against. So figuring out how to deal with implicit bias requires a better understanding of what it is. Accordingly, Part 1 of this volume introduces readers to theoretical accounts of implicit bias. These accounts serve as key reference points for the moral and political questions raised in subsequent chapters. The first chapter analyzes implicit bias from a traditional psychological perspective. How do implicit biases fit into our understanding of the mind? The second chapter broadens the discussion to ask how implicit bias may be understood in a more holistic way, namely, as residing not just "inside our minds" but in our physical bodies, habits, and social practices. The third chapter explores—and aims to answer—various skeptical arguments to the effect that implicit biases don't exist and, if they do, are not particularly helpful in understanding injustice.

Chapter 1—The Psychology of Bias: From Data to Theory

In Chapter 1, Gabbrielle Johnson introduces readers to leading psychological theories of implicit bias. According to one model, implicit biases are automatic, relatively unconscious mental associations. For example, you hear "salt," then think "pepper." You think "woman," then think "mother." According to a second model, implicit biases are unconscious beliefs.

To evaluate these models, Johnson asks what makes for a good psychological explanation, which would illuminate how the mind works. Next, Johnson explores the tools psychologists use to study the mind, and, in particular, the contrast between *direct* and *indirect* measures of people's attitudes. Direct (or *explicit*) measures ask people to report their beliefs and feelings openly (for example: "how much pain are you feeling, on a scale of 1 to 10?"). Indirect (or *implicit*) measures, such as the Implicit Association Test, instead aim to get at people's attitudes without their reporting them. Indirect measures have been pivotal for advancing our knowledge of implicit bias.

These measures reveal one of the most striking features of implicit biases, which is that they can come apart, or diverge, from our explicit beliefs and values. For example, a person might express a sincere commitment to treating members of all racial groups equally but

nevertheless demonstrate subtle racial biases on an indirect measure. How should we understand the negative 'gut reactions' or 'snap judgments' that drive performance on these measures? What kinds of theories can explain their divergence from our explicit beliefs?

Additionally, Johnson examines how emerging research upends commonsense thinking about implicit bias. Originally, implicit biases were thought to be deeply ingrained products of our upbringing that would be difficult or impossible to change. However, new evidence reveals that in some contexts, our implicit biases are surprisingly easy to change. What kinds of theories can explain why biases change when they do, and why they don't change when they don't? Answers to these abstract, theoretical questions promise serious practical payoffs, as subsequent authors in this volume explore.

Chapter 2—The Embodied Biased Mind

Implicit bias is often framed in individualistic, psychological terms. The framing is not surprising given that cognitive and social psychologists have been at the forefront in theorizing the phenomenon. Yet the dominant way of understanding bias has a problem: it creates the impression that bias exists exclusively "in the head" of individuals.

In Chapter 2, Céline Leboeuf explores a more holistic way of understanding what bias is. Drawing on the work of Maurice Merleau-Ponty, she argues that implicit biases can be thought of as *perceptual habits*. Habits are learned behaviors, which are realized by—and necessarily depend on—our bodies. If so, implicit bias would consist in bodily habits, rather than mental activity per se.

An embodied view of bias is helpful. For example, it would go a long way towards explaining why implicit biases are often experienced as automatic and beyond conscious awareness. Just as we don't have to consciously think about how to position our hand to turn off a light switch in a familiar room—our hand instinctively and thoughtlessly goes there by habit—one needn't consciously think about habitual ways of seeing others. For example, in well-known experiments, young black men are perceived as acting more aggressively than young white men, even though their behavior is identical. People who see young black men act from habit. They manifest a tendency to look at and pay attention to young black men in specific ways, and hence to engage and interact with them in particular ways. The pervasiveness of such biases is no accident. As Leboeuf notes, we pick up the habits we do in social environments, where norms and expectations about different kinds of people are communicated to us in subtle and not so subtle ways. Biases thus reflect inequalities and norms in society at large. Her analysis also reveals—through discussion of sociologist Pierre Bourdieu's work on "the habitus"—ways in which widespread biases might impact how we navigate the social world in ways that reflect and entrench group hierarchy. One upshot of the embodied approach is that we need to literally retrain ourselves and develop better habits if we aim to create a more just world (see also McHugh and Davidson, Chapter 9, "Epistemic Responsibility and Implicit Bias" and Alex Madva, Chapter 12, "Individual and Structural Interventions").

Chapter 3—Skepticism about Bias

In Chapter 3, Michael Brownstein raises and responds to six of the most prominent criticisms of implicit bias. The first, big-picture objection he considers is that the inequalities supposedly related to implicit bias either don't exist at all (for example, maybe police officers treat whites and blacks just the same), or aren't truly unfair (for example, maybe police officers arrest blacks more often than whites because blacks commit more crimes). The second overarching objection

begins by acknowledging that these inequalities are both real and really unfair, but then counters that bias has little or nothing to do with them. A third objection grants that inequalities exist, that they're unfair, and that they're partly explained by bias—but then objects that the operative biases are *explicit* rather than *implicit*.

Coming from another direction, a fourth objection—which is taken up in much greater depth in Chapters 11 and 12—is that the primary drivers of inequality are related to institutions and structures (like segregation, see above) rather than bias. Next, Brownstein steps back from these controversies about the explanatory power of implicit bias to consider a fifth criticism, which argues that implicit bias research has been largely "over-hyped." Here the objection has to do with the ways scientists and journalists have *communicated* implicit bias findings to the general public. Finally, a sixth criticism considered by Brownstein is that the basic tools social scientists have developed to study implicit bias (like the Implicit Association Test) are foundationally flawed, and that we'll need better methods to measure minds before we can fully appreciate the roles that bias and injustice play in social life.

Brownstein helpfully focuses on notable representatives of each criticism, and responds to each one by one. Along the way, he frequently grants that the critics are at least partly right, and that many outstanding challenges and unknowns remain. The upshot of these challenges, however, is not to give up on implicit bias research altogether, but to keep improving the research. In other words, it's a call to action. Researchers, scholars, and activists must redouble their efforts to understand what implicit bias is, how best to measure it, and ultimately how best to overcome it.

Part 2

What is the relationship between bias and knowledge? Part 2 explores this question. One of its key aims is to document ways in which biases can frustrate knowledge. However, it also asks whether biases could ever track truth and be reliable in some cases.

Chapter 4—Bias and Knowledge: Two Metaphors

In Chapter 4, Erin Beeghly investigates two metaphors used to characterize the relationship between bias and knowledge: bias as a kind of *fog* that surrounds us and bias as a kind of *shortcut* for forming beliefs about the world. She argues that these two metaphors point to a deep disagreement among theorists about whether biases can help us be more reliable knowers in some cases. They also help us to better understand the range of knowledge-related concerns about bias.

Examining these metaphors, Beeghly observes that biased judgments are motivated by stereotypes. One objection to biased judgments and perceptions is, therefore, that stereotypes are false and based on inadequate evidence. She argues that this objection will not always apply in cases of implicit bias due to the fact that group stereotypes—which are associated with group generalizations—may be true or accurate in some cases, at least to some degree. "Doctors wear white coats" and "Women are empathetic" are potential examples. Even so, she argues, bias may compromise knowledge in other ways. Drawing on the psychological and behavioral-economic literature on heuristics and biases, she outlines a number of ways in which biased judgments—even if grounded in accurate views of groups—may be unreliable. Her analysis suggests that a unified theory of how biases impede knowledge is unlikely. Biased judgments may frustrate knowledge in range of different ways and for different reasons.

A second theme in her analysis is the positive epistemic function of biases. If bias is like a fog, it necessarily stops us from seeing other people and the world clearly. However, if biases are shortcuts, they will sometimes be an efficient, effective mode of forming beliefs or expectations about individuals. Some biases may even track truth. Beeghly supplements this observation with what she calls "the argument from symmetry." The argument from symmetry says that biases—both implicit and explicit—reflect a more general feature of human cognition, namely, our tendency to carve up the world into categories and form expectations about individuals based on how we classify them. No one is tempted to say that category-based reasoning is *always* bad for knowledge. Differentiating things into categories—distinguishing viruses from bacterial infections, distinguishing edible mushrooms from poisonous ones, and distinguishing cats from dogs—is essential for knowing and navigating the world. So why would it always be bad to reason about *social* categories—to distinguish one racial or religious group from another? Beeghly suggests that the argument from symmetry articulates a challenge to theorists, namely, to more carefully delineate the conditions under which implicitly biased judgments are—and are not—reliable. Her analysis also raises the question whether ethical ideals could "raise the epistemic bar" when we interact with others humans, forcing us to gather more and better evidence than would otherwise be required (as Basu argues in Chapter 10, "The Specter of Normative Conflict: Does Fairness Require Inaccuracy?").

Chapter 5—Bias and Perception

In Chapter 5, Susanna Siegel explores how biases—both explicit and implicit—impact how we perceive others. She examines biased racial perception through three frames: cultural analysis, cognitive science, and epistemology. Each frame reveals something new and important about biased perception.

Cultural analysis reveals "what it is like" to see others in a racialized way or to have such perceptions foisted upon you. Siegel reports the testimony of George Yancy, a black philosopher who describes walking through a department store. Yancy writes, "I feel that in their eyes [that is, in the eyes of white employees and shoppers] I am this indistinguishable, amorphous, black seething mass, a token of danger, a threat, a rapist, a criminal, a burden." Siegel also describes the hypothetical example of Whit, a white person who grows up in an all-white town who experiences men of color as suspicious.

Turning to cognitive science, Siegel examines how empirical research sheds light on biased cognition and perception. In one study, research participants tended to misclassify benign objects like pliers as guns when black men were holding them. How do such mistakes arise? Siegel canvasses seven options. One option is that participants literally see the pliers as a gun, hence are subject to a visual illusion due to their biases. Another option is that participants correctly see the pliers as pliers but mistakenly push the button that indicates that they have seen a gun due to bias. In that case, their perception is accurate, but they act in biased way by compulsion. As she explores different kinds of mistakes that could manifest in biased perception and action, Siegel returns to an issue raised in Part 1: are biases "mere associations" or beliefs? She argues that modeling biases as akin to ordinary beliefs is more plausible, given the empirical research.

Finally, Siegel deploys the lens of epistemology. If biases take the form of beliefs, we can evaluate them in terms of whether they are justified or unjustified, and whether they advance or undermine our knowledge of the social world. She then raises a worry about the nature of perception, namely, that it can be "hijacked" by ill-founded, inaccurate outlooks on others. When

a person's perception is hijacked by racist stereotypes, she notes, their experiences will seem reasonable to them, as demonstrated by cultural analysis. For example, the shoppers who look at Yancy with suspicion will see their worries about him as justified. To them, he really does look like he is up to no good. However, such visual experiences are unreasonable. People should, in these cases, not believe what they see. Her conclusion cuts against the common view that visual perception is one of the most reliable forms of knowledge—which has profound implications for policing and law.

Chapter 6—Epistemic Injustice and Implicit Bias

Each one of us has the ability to produce knowledge and contribute to collective inquiry. These knowledge-producing abilities are part of what gives our life meaning and makes it valuable. In Chapter 6, Jules Holroyd and Katherine Puddifoot use this observation as a springboard. Here is what they say: because our knowledge-generating abilities are connected to our moral worth as individuals, we can wrong other people by treating them in ways that are disrespectful of their status as knowers. Such wrongs are called "epistemic injustices."

Using the film *Hidden Figures*, which charts the key contributions and unjust experiences of black women in NASA's early space program, Holroyd and Puddifoot discuss five forms that bias-driven epistemic injustice can take. *Testimonial injustice*, for example, occurs when people are believed less than they would be otherwise due to pernicious group stereotypes (e.g., when the police don't believe you because you're black). *Epistemic appropriation* occurs when people do not get adequate credit for the ideas they produce due to unfair power dynamics, which often involve bias (e.g., "he-peating," when a woman puts forward an idea at a meeting and no one responds, but then a man repeats the same idea later and gets credit for it). *Epistemic exploitation* occurs when members of marginalized groups bear the burden of educating members of dominant groups about the injustices they face (e.g., when white people expect their black acquaintances to explain what's wrong with dressing up in blackface, rather than just looking up the answers themselves). According to Holroyd and Puddifoot, injustices such as these are ever-present in social life. This is to be expected, they argue: knowledge and power are inseparable. The project of fighting epistemic injustice must therefore be a collective enterprise.

Part 3

Building on Part 1's discussion of the psychological nature of implicit bias, and Part 2's discussion about implicit bias and knowledge, Part 3 turns to questions of *morality and justice*. What are the practical, ethical, and political implications of implicit bias?

Chapter 7: Stereotype Threat, Identity, and the Disruption of Habit

Implicit biases don't just affect how we judge other people; they also affect how we see and judge ourselves. In this way, the phenomenon of implicit bias is closely related to *stereotype threat*. Roughly, stereotype threat occurs when being reminded of one's social identity and the stereotypes associated with it (such as gender and racial stereotypes) leads to anxiety, alienation, and underperformance.

In Chapter 7, Nathifa Greene investigates this phenomenon and introduces readers to different ways of understanding its importance. According to the standard view, stereotype threat occurs when a person fears that she will vindicate negative group stereotypes. For example, a woman might do worse on a math exam if her gender is made salient to her or if she is reminded of negative stereotypes like "women are bad at math." In social psychology, researchers have

attempted to document the effects of stereotype threat and hypothesize about what happens inside the mind of someone who experiences it. Greene points out numerous problems with the standard approach. First, empirical studies of stereotype threat have not always been replicated, and serious doubts exist about the robustness of empirical findings. Second, the standard view makes it seems as if people who experience stereotype threat are irrational: they shouldn't doubt their abilities or themselves because they are fully capable, but they do. Third, the view implies that people are simply stereotyping themselves; if so, the standard account further harms victims of stereotype threat by suggesting that they are the root of the problem.

Greene suggests an alternative view. In particular, she argues that stereotype threat primarily consists in a form of *disruption*, when an individual cannot just "be" in the world with one's skills and habits, but gets knocked out of the "flow." In this way, stereotype threat is similar to the experience of "choking" that athletic and artistic performers can experience. Drawing on the work of W.E.B. Du Bois and Frantz Fanon, Greene persuasively argues that knowledge of the phenomena related to stereotype threat existed long before social psychologists began to study it in the 1990s. Further, she suggests that the perspective of cognitive science sometimes hides rather than reveals the true nature of the phenomenon. From a first-person perspective, we see stereotype threat is not irrational: it involves correctly perceiving that others—often, those in more socially privileged groups—are prone to think badly of you due to group stereotypes. Moreover, people suffer stereotype threat not because they stereotype themselves out of insecurity or anxiety but because negative stereotypes are foisted upon them in everyday social environments. From Greene's analysis, a vision of a just society emerges in which people are able to seamlessly and safely navigate their world, and inhabit their bodies, without the imposition of others' harmful implicit and explicit biases.

Chapter 8—Moral Responsibility for Implicit Biases: Examining our Options

Are we responsible for our knowledge and how we act on biases in everyday settings? Is it ever appropriate to blame or hold individuals accountable when their actions are subtly or not-so-subtly influenced by implicit biases? Answering these questions requires that we first step back to consider what makes people morally responsible *in general*. Why is anybody ever responsible for anything?

In Chapter 8, Noel Dominguez surveys leading theories about the nature and necessary conditions for moral responsibility. He asks which verdicts each of these theories delivers about responsibility for implicit bias. Answering this question is difficult, he argues, because much about implicit bias remains unknown and our scientific "best guesses" keep evolving (as discussed in Part 1 of the volume).

One leading theory of responsibility argues that we can only be held responsible for actions that we intentionally choose to do, that is, actions within our *control*. If I bump into you because somebody pushed me, then that's out of my control, and it seems inappropriate to hold it against me. But if I bump into you because, well, that's what I chose to do, then maybe I deserve your anger. Then the question for implicit bias becomes: are we enough in control of our biases to be responsible for them? Perhaps we can't control them *directly*, in the moment, but maybe we can control them *indirectly*, by cultivating the sorts of long-term habits or social policies (such as by grading papers and reviewing job applications anonymously). Indirect strategies likely won't work against all biases all the time, however.

A second leading theory states that control is not necessary for responsibility, because what really matters is whether our actions reflect "deep" facts about our character traits and values. Consider one example introduced by Angela Smith (2005). If I forget my best friend's birthday, then plausibly my *failure to remember* was not the result of any conscious, controlled choice—and yet it still seems appropriate for my friend to resent me for forgetting. Why might that be? Maybe my failure to remember was a result of not *caring* about my friend as much as I should, and so reflects a "deep" fact about who I am and what I value. Then the question for implicit bias becomes: do implicit biases reflect deep facts about who we are as individuals, or are they just superficial features of our minds, or generic biases that most folks in our culture absorb? Dominguez considers arguments on both sides. Ultimately, he thinks that it would be unfair to hold people responsible for every aspect of their deepest self.

Dominguez concludes by arguing that research on implicit bias may force us to *revise or revolutionize* our understanding of moral responsibility altogether. To know whether such revisions are justified, he says, we must think harder about what we want a theory of moral responsibility to do.

Chapter 9—Epistemic Responsibility and Implicit Bias

A topic of special importance when it comes to responsibility and implicit bias is responsibility for *knowledge*. Are there strategies for becoming more responsible and respectful knowers? How might we work together to reduce the negative effects of bias on what we see and believe, as well as the wrongs associated with biases? In Chapter 9, Nancy Arden McHugh and Lacey J. Davidson explore these questions. Like Holroyd and Puddifoot in Chapter 6 ("Epistemic Injustice and Implicit Bias"), they argue that adequately answering them requires thinking about responsibility as having both *individual* and *collective* dimensions.

Their article begins with a discussion of moral responsibility for bias. They argue that typical discussions of responsibility—such as those discussed by Dominguez in Chapter 7, "Moral Responsibility for Implicit Biases: Examining Our Options"—tend to think of responsibility exclusively as an individual matter. They argue that individualistic approaches lead to puzzles that misleadingly suggest that we should not be held responsible for our biases and that implicit biases don't belong to us. These puzzles disappear, they contend, if we recognize the collective dimensions of responsibility. They thus introduce the concept of *epistemic responsibility*, which they believe better tracks the social and collective aspects of responsibility. Epistemic responsibility, they explain, "is a set of habits or practices of the mind that people develop through the cultivation of some basic epistemic virtues, such as open-mindedness, epistemic humility, and diligence that help knowers engage in seeking information about themselves, others, and the world that they inhabit."

Realizing these virtues requires putting oneself in a larger social frame. Note the centrality of habits and practices (as emphasized in Leboeuf, Chapter 2, "The Embodied Biased Mind" and Greene, Chapter 7, "Stereotype Threat, Identity, and the Disruption of Habit"). Habits and practices are social in that they are acquired in the context of communities. Knowers always exist in a time and place, the view goes, and we acquire the habits we do partially in virtue of how we are raised. It follows that we need to work together to make our world one in which epistemic virtues like open-mindedness, creativity, and self-reflection can flourish. In a world that is highly unjust, this is a serious challenge. Nonetheless, McHugh and Davidson offer a way forward. Outlining four promising strategies for combatting bias-related epistemic vices, they argue that we can change our world—and ourselves—for the better. "World traveling," they

note, drawing on the work of María Lugones (1987), requires actively seeking out social situations that are outside your comfort zone and interacting with people who are different from you in order to challenge your own way of thinking. They also consider the strengths and weaknesses of practices like "calling in" and "calling out." When you call someone in, you confront that person about their biased or otherwise insensitive behavior, but you do so in a way that aims to build a stronger relationship with them and alerts them to the harm they have caused. The aim is restorative. They also consider whether punitive "call outs" are always counterproductive, or might sometimes have their place. Recommendations like this—and others—are explored in greater depth in Chapter 12 (Madva, "Individual and Structural Interventions").

Chapter 10—The Specter of Normative Conflict: Does Fairness Require Inaccuracy?

A further question surrounding responsibility, knowledge, and bias is this: is it always morally wrong to rely on stereotypes when making judgments about other people? What if the stereotypes are sometimes *accurate* (as Beeghly suggests in Chapter 4, "Bias and Knowledge: Two Metaphors")? Does judging people fairly mean that you must ignore what is most likely true about them given your evidence? While Chapters 8 and 9 (Dominguez, "Moral Responsibility for Implicit Biases: Examining Our Options"; McHugh and Davidson, "Epistemic Responsibility and Implicit Bias") focus on responsibility and action, Chapter 10 turns to rationality and belief, exploring the relationship between *knowledge* and *justice*. Rima Basu asks whether research on social bias pits *fairness* against *accuracy*.

One intuitive thought—which has been endorsed by Tamar Szabó Gendler—is that we are faced with a tragic dilemma in everyday situations because we live in an unjust world. Our evidence tells us that we should have certain beliefs about people, but ethical norms forbid it. Suppose, for instance, you are eating at a restaurant where ninety-percent of employees are people of color and almost all restaurant patrons are white. If you see a person of color at the other end of the room, the statistical evidence apparently suggests that this person is an employee. On one hand, it seems like you should believe what the evidence tells you. On the other hand, it is wrong to stereotype someone based on race. Forming beliefs about someone based on their perceived race is unfair after all, and it can be experienced as harmful. So you seem to face a hard choice. Ignore your evidence. Or, believe your evidence but judge someone unjustly.

As Basu explains, there are good reasons to question whether cases like these truly present irresolvable dilemmas, and whether accuracy and fairness are diametrically opposed. Basu thus sketches and evaluates a range of alternative views. One theory, for instance, is that ethical norms are inherently superior to norms related to knowledge and, thus, take priority over them. After raising challenges for this theory and others, Basu suggests that ethical and epistemic norms ultimately work together rather than in opposition. In cases where the moral stakes are high, Basu argues, it is never appropriate to judge someone based merely on statistics or stereotypes. You always need more and better information, tailored to the individual. Her view suggests an exciting possibility, namely, that it is often—if not always—ethically and epistemically wrong to judge people based solely on probabilistic evidence, and even on reliable generalizations.

Chapter 11— Explaining Injustice: Structural Analysis, Bias, and Individuals

Several contributors to this volume stress that implicit bias and social injustice are neither best understood nor best overcome solely at the individual level. The critique is persuasive: the specific actions of implicitly and explicitly biased individuals are not the whole story about how the most egregious patterns of injustice and inequality arise and persist. Nor are they the only place to look when we think about potential ways forward. In Chapter 11, Saray Ayala-López and Erin Beeghly offer an in-depth analysis of these and related points.

The chapter begins with two hypothetical examples of social injustice. Individualistic approaches understand injustices such as these as the result of individuals' preferences, beliefs, and choices. For example, they explain racial injustice as the result of individuals acting on racial stereotypes and prejudices. Structural approaches, in contrast, appeal to the *circumstances* in which individuals make their decisions. For example, they explain social injustice in terms of beyond-the-individual features, including group dynamics, laws, institutions, city layouts, and social norms. Often these two approaches are seen as competitors. Framing them as competitors, Ayala-López and Beeghly argue, suggests that only one approach can win and that the loser offers worse explanations of injustice.

Before they evaluate this claim, Ayala-López and Beeghly ask whether individualistic and structural approaches are as different as people often think. They argue that the answer is no. The best accounts of implicit and explicit bias, for example, see individual psychology as responsive to and as an expression of social structures. Hence explanations of injustice that cite individuals' biases need not ignore the existence of social structures. Indeed some bias-focused explanations suggest that structural factors provide the deepest explanations of why social injustices occur and persist. Likewise, structural accounts can—and often do—cite the actions of individuals when they explain injustice. If so the two approaches are better seen not as diametrically opposed but as making claims about which explanatory factors should be *prioritized*.

Working with this more nuanced picture, Ayala-López and Beeghly step back and explore criteria for comparing and evaluating the two approaches. Does one approach better identify what's ethically troubling about injustice? Does one have a better story about why social injustices occur? And, does one approach provide superior strategies for changing our world for the better? Ayala-López and Beeghly argue that no one approach has the upper hand with respect to answering these questions in any and all contexts. If they are right, both approaches are needed to adequately understand and address injustice. They contend, further, that the two approaches can work together—synergistically—to produce deeper explanations of social injustice.

Chapter 12: Individual and Structural Interventions

Given all that we have learned about bias and injustice, what can we do—and what should we do—to fight back? In our volume's final chapter, Alex Madva builds on the philosophical, psychological, and ethical insights of earlier chapters to make the case for a set of scientifically-tested *debiasing* interventions—that is, interventions to combat implicit (and explicit) bias as well as promote a fairer world.

Madva begins by noting a few of the leading obstacles to meaningful social change, stressing in particular *the gaps in what we know*, both about which concrete goals we should aim

for and how best to pursue them. He illustrates our knowledge limitations by focusing on two case studies of persistent injustice that have proven difficult to overcome, namely, the gender pay gap and the challenges of reentering society for people getting out of prison. Madva shows how the seemingly obvious ways to fix these injustices sometimes do nothing, and sometimes make matters worse. He next draws several lessons from these case studies. One is that we must adopt an *experimental mindset*: we have to test out different strategies and see how they go, then go back to the drawing board, revise our strategies, and test them again. Another lesson is that, given how multifaceted these problems are, we have to adopt an equally multifaceted approach to solving them. This means, among other things, that we will need to make changes both as individuals (how can I do better?) and as communities (how can we do better?).

With this ground cleared, Madva then turns to recommending concrete strategies. These range from small-scale, daily-life debiasing tricks, like perspective-taking, to large-scale societal transformations. Each step of the way, Madva teases out various moral, political, and strategic questions for these interventions, and points the reader to the outstanding unknowns about them. In doing so, he highlights the many gaps in our knowledge that future scientists, activists, artists, and frankly *leaders of all kinds* will have to fill.

None of the large-scale transformations that Madva, and other authors in this volume, consider will be possible with the snap of a finger. They'll only be brought about through a correspondingly large social movement. This volume is not a textbook for how to activate and sustain that movement. Yet some of the challenges caused by implicit and explicit bias in social life more generally will also arise in the context of social movements (Cortland et al., 2017; Jost et al., 2017; van Zomeren, 2013). When we must build coalitions across boundaries of class, race, gender, sexuality, ability, and other dimensions of difference, we will be faced with biased people with whom we must work if we want our movement to succeed, and we will inevitably be forced to reckon with biases in ourselves. In such contexts, individuals have to find a way to build solidarity in the face of bias. There are no easy answers here. If nothing else, carefully thinking through the hard questions discussed in this volume provides a step in the right direction. We hope it motivates readers to do their small part, in their own corner of the world, and to be ready to join into that movement when it arises. *Hint*: the time is now!

References

Alexander, M., 2012. The New Jim Crow: Mass Incarceration in the Age of Colorblindness. The New Press, New York.

Anderson, E., 2010. The Imperative of Integration, Reprint edition. ed. Princeton University Press, Princeton, N.J.

Ayala-López, S., 2018. A Structural Explanation of Injustice in Conversations: It's about Norms. Pacific Philosophical Quarterly 99, 726–748. https://doi.org/10.1111/papq.12244

Blanton, H., Ikizer, E.G., 2019. Elegant Science Narratives and Unintended Influences: An Agenda for the Science of Science Communication: Unintended Influence. Social Issues and Policy Review 13, 154–181. https://doi.org/10.1111/sipr.12055

Brownstein, M., 2018. The Implicit Mind: Cognitive Architecture, the Self, and Ethics. Oxford University Press.

Cable, D., 2013. The Racial Dot Map | Weldon Cooper Center for Public Service [WWW Document]. URL https://demographics.coopercenter.org/racial-dot-map (accessed 8.14.19).

Cortland, C.I., Craig, M.A., Shapiro, J.R., Richeson, J.A., Neel, R., Goldstein, N.J., 2017. Solidarity through shared disadvantage: Highlighting shared experiences of discrimination improves relations between stigmatized groups. Journal of Personality and Social Psychology 113, 547–567. https://doi.org/10.1037/pspi0000100

Crandall, C.S., Eshleman, A., O'Brien, L., 2002. Social norms and the expression and suppression of prejudice: The struggle for internalization. Journal of Personality and Social Psychology 82, 359–378. https://doi.org/10.1037//0022-3514.82.3.359

De Houwer, J., In press. Moving beyond the distinction between System 1 and System 2: Conditioning, implicit evaluation, and habitual responding might also be mediated by relational knowledge. Experimental Psychology.

Dunham, Y., Baron, A.S., Banaji, M.R., 2008. The development of implicit intergroup cognition. Trends in Cognitive Sciences 12, 248–253. https://doi.org/10.1016/j.tics.2008.04.006

Enos, R.D., 2017. The space between us: social geography and politics. Cambridge University Press.

Feagin, J.R., Sikes, M.P., 1995. Living with Racism: The Black Middle-Class Experience, Reprint edition. ed. Beacon Press, Boston.

Gawronski, B., Bodenhausen, G.V., 2006. Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. Psychological Bulletin 132, 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Haslanger, S., 2015. Social Structure, Narrative, and Explanation. Canadian Journal of Philosophy 45, 1–15.

Hermanson, S., 2017. Implicit Bias, Stereotype Threat, and Political Correctness in Philosophy. Philosophies 2, 12. https://doi.org/10.3390/philosophies2020012

Johnson, J., 2016. Two days after the debate, Trump responds to Clinton's comment on implicit bias. Washington Post.

Jost, J.T., Becker, J., Osborne, D., Badaan, V., 2017. Missing in (Collective) Action: Ideology, System Justification, and the Motivational Antecedents of Two Types of Protest Behavior. Current Directions in Psychological Science 26, 99–108. https://doi.org/10.1177/0963721417690633

King Jr., M.L., 1963. Letter From a Birmingham Jail [WWW Document]. URL https://kinginstitute.stanford.edu/king-papers/documents/letter-birmingham-jail (accessed 1.6.19).

Lauer, H., 2019. Implicitly Racist Epistemology: recent philosophical appeals to the neurophysiology of tacit prejudice. Angelaki 24, 34–47. https://doi.org/10.1080/0969725X.2019.1574076

Lee, K.M., Lindquist, K.A., Payne, B.K., 2017. Constructing Bias: Conceptualization Breaks the Link Between Implicit Bias and Fear of Black Americans. Emotion. https://doi.org/10.1037/emo0000347

Lopez, G., 2017. The past year of research has made it very clear: Trump won because of racial resentment [WWW Document]. Vox. URL https://www.vox.com/identities/2017/12/15/16781222/trump-racism-economic-anxiety-study (accessed 12.17.17).

Lugones, M., 1987. Playfulness, "World"-Travelling, and Loving Perception. Hypatia 2, 3–19.

Mac Donald, H., 2017. The False 'Science' of Implicit Bias. Wall Street Journal.

Mac Donald, H., 2016. The War on Cops: How the New Attack on Law and Order Makes Everyone Less Safe, First Edition edition. ed. Encounter Books, New York.

Madva, A., 2019. Social Psychology, Phenomenology, and the Indeterminate Content of Unreflective Racial Bias, in: Lee, E.S. (Ed.), Race as Phenomena: Between Phenomenology and Philosophy of Race. Rowman & Littlefield International, Lanham, pp. 87–106.
Ngo, H., 2019. The Habits of Racism: A Phenomenology of Racism and Racialized Embodiment. Lexington Books, S.l.
Oswald, F.L., Mitchell, G., Blanton, H., Jaccard, J., Tetlock, P.E., 2013. Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. Journal of Personality and Social Psychology 105, 171–192. https://doi.org/10.1037/a0032734
Payne, B.K., Krosnick, J.A., Pasek, J., Lelkes, Y., Akhtar, O., Tompson, T., 2010. Implicit and explicit prejudice in the 2008 American presidential election. Journal of Experimental Social Psychology 46, 367–374. https://doi.org/10.1016/j.jesp.2009.11.001
Payne, B.K., Vuletich, H.A., 2017. Policy Insights From Advances in Implicit Bias Research. Policy Insights from the Behavioral and Brain Sciences 2372732217746190. https://doi.org/10.1177/2372732217746190
Payne, B.K., Vuletich, H.A., Brown-Iannuzzi, J.L., 2019. Historical roots of implicit bias in slavery. PNAS 201818816. https://doi.org/10.1073/pnas.1818816116
Pettigrew, T.F., 1998. Intergroup Contact Theory. Annual Review of Psychology 49, 65–85. https://doi.org/10.1146/annurev.psych.49.1.65
Rothstein, R., 2018. The Color of Law: A Forgotten History of How Our Government Segregated America, 1 edition. ed. Liveright, New York London.
Shelby, T., 2016. Dark ghettos: injustice, dissent, and reform. The Belknap Press of Harvard University Press, Cambridge, Massachusetts.
Singal, J., 2017. Psychology's Favorite Tool for Measuring Racism Isn't Up to the Job. Science of Us, New York Magazine.
Smith, A.M., 2005. Responsibility for Attitudes: Activity and Passivity in Mental Life. Ethics 115, 236–271. https://doi.org/10.1086/426957
Van Den Bergh, B., Heuvinck, N., Schellekens, G.A.C., Vermeir, I., 2016. Altering Speed of Locomotion. J Consum Res 43, 407–428. https://doi.org/10.1093/jcr/ucw031
van Zomeren, M., 2013. Four Core Social-Psychological Motivations to Undertake Collective Action. Social and Personality Psychology Compass 7, 378–388.
White, M.H., Crandall, C.S., 2017. Freedom of racist speech: Ego and expressive threats. Journal of Personality and Social Psychology 113, 413–429. https://doi.org/10.1037/pspi0000095