Michael Brownstein and Jennifer Saul, editors
Implicit Bias and Philosophy: Metaphysics and Epistemology, volume 1
Oxford: Oxford University Press, 2016 (ISBN: 9780198713241)

Saray Ayala-López

Saray Ayala-López is an assistant professor of philosophy at California State University, Sacramento. They have previously worked at San Francisco State University, Carlos III University of Madrid, and Autonomous University of Barcelona. Two extreme emotions brought Ayala to philosophy: a passion for science and understanding, and a deep disappointment with social injustice. Their current research applies conceptual tools from the philosophy of science (especially cognitive science), language, and mind, to a variety of questions in social and feminist philosophy, for example, explanations of social injustice, the dynamics of conversations, the use of sex categories, the metaphysics and epistemology of sexual orientation. They recently started a platform for foreigners in academia, part of a broader project to give visibility to the difficulties that "alien scholars" face. They publish as both Saray Ayala and Saray Ayala-López.

******

We all have heard of implicit bias, and many of us have appealed to implicit bias as part of an explanation or a diagnosis, in both ordinary conversations and also in academic papers. But as it happens with trending topics, many of us have perhaps rushed into adopting implicit bias talk uncritically without knowing much about it. This collection edited by Michael Brownstein and Jennifer Saul, the first volume of two, provides a fine philosophical toolkit on implicit bias. It sheds light on some of the complex issues surrounding implicit bias, and it invites us to ask more and better questions about this morally charged cognitive phenomenon.

The book starts with a brief introduction to implicit attitudes, which includes a description of the different ways they have been characterized in the empirical literature, and of different measures of implicit attitudes, including the well-known Implicit Association Test. The remaining chapters are divided into two groups. The first group addresses questions about the metaphysics of implicit bias and stereotype threat, for example, what sort of thing are implicit biases? How unified are they? What are the mechanisms underlying implicit bias and stereotype threat? The second group addresses epistemological questions, and it pays attention to the epistemic costs of implicit bias and stereotype threat, for example, how does implicit bias affect the profession of philosophy and our scientific endeavors? Does implicit bias give rise to a new breed of skepticism, as Saul diagnosed (Saul 2013)? Does it lead to a dilemma between our epistemic and our moral duties? What are the epistemic costs of stereotype threat?

Since space constraints prevent me from explaining each of these chapters individually, I focus on two chapters as they address two critical questions. The first question is whether empirical results and theoretical generalizations about implicit bias all pertain to a single, unified phenomenon, or rather a variety of them. Second, if our evaluations of others (for example, their CVs, their credibility) are influenced by factors that we ourselves recognize to be irrelevant (for

https://doi.org/10.1017/S2753906700002242 Published online by Cambridge University Press

example, their perceived social identity), what does that say about the reliability of our epistemic practices? And what are the consequences for our epistemic agency?

Jules Holroyd and Joseph Sweetman's chapter focuses on the first question. They express concerns about the oversimplification found in many works on implicit bias, where implicit biases are referred to as whatever unconscious processes are influencing our judgments and perceptions of others. The authors acknowledge that this overarching notion is useful for particular purposes, for example, when the focus is on the general effects of implicit bias. Despite these advantages, they point out two dangers in how implicit bias appears in the philosophical discourse, both having to do with taking implicit bias as a single, homogeneous phenomenon. First is the danger of unwarranted generalizations. This might happen if we use the term *implicit bias* too loosely, to refer to a range of different processes. Second is the danger of putting forward misleading or ineffective normative recommendations. This happens if we ignore the differences among implicit biases, and so fail to appreciate the ways in which these differences are relevant for designing interventions aimed at reducing or eliminating them.

The heterogeneity of implicit bias is evident in how it stands in relationship to explicit beliefs. Some studies suggest independence between implicit and explicit beliefs (for example, Banaji and Hardin 1996), whereas others suggest important dependence between them (for example, Devine et al. 2002). We also see heterogeneity in the different behavioral predictions that implicit biases make. In a study by David Amodio and Patricia Devine, although strong associations between racial categories and stereotypical traits predicted competence judgments consistent with stereotypes, it did not predict negative affective behavior (Amodio and Devine 2006). Critics of implicit bias might see contradictory results here, and proof that research on implicit bias is not providing interesting, reliable results. A less radical critique would be that observed heterogeneity is due to deficits in experimental designs: as we improve the latter, the reasoning goes, we should be able to finally get at what we are looking for, that is, *the* phenomenon of implicit bias. Holroyd and Sweetman argue that the best way to explain observed heterogeneity is to take it as a reflection of what is really going on with implicit bias: there is no single phenomenon, and so the goal in empirical research on implicit bias should not be to reduce complexity (and apparent contradictions) in our results, but to gain a better understanding of it.

The authors' analysis of the affective vs. semantic distinction in empirical psychology is a good illustration of the important role philosophy plays in empirical research on implicit bias. The semantic vs. affective distinction is supposed to distinguish between implicit associations that have an affective valence, and those that are semantically related. Holroyd and Sweetman problematize this distinction in two ways: as it is used in a study by Amodio and Devine (Amodio and Devine 2006), and more generally as a dimension of heterogeneity in implicit associations. This analysis is a great illustration of how research on implicit bias needs a full room of armchairs to fine-tune our categorizations and adjust and readjust our theoretical frameworks.

If implicit bias is a heterogeneous set of phenomena, a question arises: How heterogeneous should interventions to eliminate bias be? Holroyd and Sweetman recommend that intervention strategies need to be sensitive to the differences among implicit biases; otherwise they risk being

https://doi.org/10.1017/S2753906700002242 Published online by Cambridge University Press

ineffective. A further question the authors do not address is whether given the heterogeneity of implicit biases, it is at all feasible to attempt to eliminate bias in people. One worry here is that this heterogeneity, combined with individual variability, might render futile any general recommendations to eliminate bias. I would like to connect this question to the broader debate on the effectiveness of interventions against social injustice. This debate is polarized between structural approaches (for example, Anderson 2010; Ayala and Vasilyeva 2015; Haslanger 2015) and individualist approaches (for example, Fricker 2007; Saul 2013). Whereas the latter favor the individual as the locus of intervention (for example, people's implicit attitudes), the structural approach emphasizes intervention on social structures and institutions. The surge in studies on implicit bias has motivated worries about paying too much attention to how individual minds contribute to injustice, and perhaps ignoring, or not giving enough attention to, the structural factors.[1]

In a recent paper, Alex Madva argues against those who prioritize structural interventions (Madva 2016), and in particular, he warns against the following move, which he suspects might be the basis of some anti-individualist approaches: taking the lack of effectiveness of a particular strategy to eliminate bias as a reason to conclude that interventions in individuals' minds are ineffective, and therefore as a reason to favor a structural approach. Holroyd and Sweetman's paper strengthens Madva's point. If implicit bias is not a unified, single phenomenon, we should not expect uniformity in how it responds to interventions.[2] Therefore, failure of an intervention targeting this or that type of implicit association is not a definite guide to the effectiveness of bias-elimination efforts in general. If taken seriously, heterogeneity seems to imply that interventions addressed toward changing people's attitudes need to be tailored to specific associations. This is an important claim that researchers on both sides, structural and individualist, need to consider. Thus Holroyd and Sweetman's chapter is especially relevant for the debate between structural and individualist approaches to social justice.

Louise Antony addresses the second question we mentioned above, that is, what are the epistemological consequences of implicit bias? Whereas Holroyd and Sweetman focus on differences among types of implicit biases, Antony looks at what biases have in common. Antony starts with bad news: since implicit biases affect our decisions in covert ways, there are reasons to worry about our epistemic stance. Saul introduced this skeptical diagnosis (Saul 2013), which Antony calls *Saulish skepticism*, and that she characterizes as a more worrying breed of skepticism compared to traditional skepticism. Whereas traditional skepticism is about possibility, that is, the possibility that some of our epistemic practices might not be reliable, Saulish skepticism is about actuality: Research on implicit bias indicates that some of our epistemic practices *are not* reliable. In the Saulish picture, the existence of implicit bias gives us reasons to doubt our epistemic agency. Antony proposes a way to escape this skepticism, and that is not by trying to eliminate (specific, negative) biases, but by embracing them (in a general

[1] See Ayala 2016 for a discussion of several concerns with regard to intervening in people's minds in a quest for a just society.

[2] See Ayala 2017, Haslanger 2017, and Saul 2017 for further comments and questions on Madva's argument. These contributions are part of a recent symposium on *The Brains Blog* (Brains Blog 2017).

https://doi.org/10.1017/S2753906700002242 Published online by Cambridge University Press

sense). The strategy is to take a naturalist approach to cognition. This will reveal at least two things. First, some of our epistemic norms may not be aligned with how our cognitive systems actually work, and we need to make peace with that. Second, biases are not (inherently) bad.

Antony reminds us that biases are necessary for getting to know the world around us. Biases help us narrow down the range of possible hypotheses to consider when we are trying to figure something out. Given the limitations of our cognitive machinery, the always-limited information we get from the world, and the fact that our everyday decisions are often made under time pressure, it is good to have some mechanism that inclines us in some direction. We need shortcuts in order to manage the vastness of the world with our limited capacities. The Adaptive Behavior and Cognition group at the Max Planck Institute for Human Development has amply explored the benefits of cognitive shortcuts Antony is talking about. As the title of one of the group's books reads, these shortcuts are *Simple Heuristics That Make Us Smart* (Gigerenzer et al. 1999). Heuristics are processes that ignore some or most of the available information and that do not aspire to an optimal outcome, but to one that is good enough. Researchers like Gerd Gigerenzer defend heuristics as not second-best, but as processes that can actually have better results than a slow and cognitively costly maximization calculus. Some of these shortcuts raise no moral concern. For example, the gaze heuristic works well when we want to catch a ball and we have no time, and probably no capacity either, to calculate its trajectory. Things get tricky, however, when we look at heuristics in the social world, and in particular in morally relevant situations. Heuristics like "go with the default" or "imitate your peers," commonly used in human decision-making, maintain the status quo, and even when they deliver benefits for the agent, they might do so with high costs for others.[3]

Antony does not argue for the benefits of all cognitive shortcuts, and in particular, she is not trying to persuade us about the hidden advantages of the implicit biases against social groups that motivated this volume in the first place. What she is arguing is that acknowledging the positive role that biases have in human cognition, and learning more about this role, prepares us to better fight the pernicious biases that conflict with our concerns for justice. It also helps us escape the skeptical predicament: by understanding the workings of biases in how we learn and make decisions, we can resist the skeptical threat. Antony gives us some guidance on how to do this. Testing the ecological validity of certain biases, and understanding why they are or are not ecologically valid in our societies, is a way of getting some control over our pernicious biases. We can do this by testing the reliability of the associations that those biases exploit (associations between a target property and what we take to be a marker of that property). Some of those associations are simply not there. Revealing this, however, is not enough to make those biases disappear. Instead, Antony proposes that we must retune the markers, which requires taking conscious steps to compensate for the distorting effects. However, this strategy against biases requires the cultivation of epistemic virtues, and it has been criticized as unattainable (for example, by Sherman 2016, in relation to biased attributions of credibility).

Interestingly, many of those associations are reliable. As Sally Haslanger recently reminded us, many morally problematic attitudes get the social reality right: "Women actually are more submissive than men; we are better caregivers than men; we are better at multi-tasking too"

[3] See Gigerenzer 2010 for a defense of heuristics in the moral domain as well.

https://doi.org/10.1017/S2753906700002242 Published online by Cambridge University Press

(Haslanger 2017, 3). In these cases we need to inquire into why this is so; that is, we need to ask what the mechanism is that connects the marker and the target property. After the observation quoted above, Haslanger continues: "This is not to say that this is true 'by women's nature' but because of the social history of gender" (3). In Antony's words, this inquiry into the mechanisms will reveal that, for many of the biases operating in social interactions, these connections are the result of "a pattern of social inequalities that we can and ought to change" (Antony, 185). Antony's call for intervention in these cases aligns with Gigerenzer's (Gigerenzer 2010). The idea is to intervene on the environment that biases exploit, and make it so that relying on heuristics does not lead to bad judgments, discrimination, and injustice. "To improve moral behavior towards a given end, changing environments can be a more successful policy than trying to change beliefs or inner virtues" (Gigerenzer 2010, 530). It seems also to align in an important sense with the recommendation that Bryce Huebner put forward in his own contribution to this book: "If we want to *overcome* implicit bias, and if we want to become the sorts of agents who are not dominated by reactions that we cannot reflectively avow, we must engage in collective prefigurative practices designed to create a world where our reflexive reactions are already calibrated against our reflectively held goals and values" (73; emphasis in the original).This type of intervention acknowledges that people rely on shortcuts; that's how our minds cope with a complex world. Instead of fixing our limited minds, let's fix our corrupted and unequal society. This predicament ties back to our previous discussion about structural vs. individualist (or psychological) interventions against social injustice. Acknowledging both the generally beneficial role biases play in human cognition, and the systematic inequalities that permeate our societies, seems to invite a structural approach to intervention.

Although I have addressed only two articles in detail, I hope to have shown just how relevant this book is for current discussions within social philosophy, epistemology, metaphysics, and moral psychology. Given the strong reliance on empirical literature, for example, providing alternative interpretations of existent empirical studies (as in Carole Lee's contribution), presenting empirical results themselves (as in Laura Di Bella, Eleanor Miles, and Jennifer Saul's contribution), the chapters in this book are not only interesting to philosophers but also to empirical researchers. Another strength of this book is the variety of questions it covers, ranging from a new understanding of stereotype threat (Ron Mallon's chapter) to expanding on the epistemic consequences of this phenomenon (Stacey Goguen's chapter); from the proposal that implicit biases are not mental states (Edouard Machery's chapter) to a prescription on how to pursue the ethical ideal of being unprejudiced without incurring epistemic costs (Alex Madva's chapter).

There is, however, a weakness in this book, and that is its lack of engagement with the criticisms that research on implicit bias has received (except for perhaps Machery's contribution). Many of those criticisms are directed toward the Implicit Association Test (the most recent one being Lai et al. 2017),[4] and so they might not be seen as interesting to discuss in a volume that aims at broadening the scope of this topic. And yet these criticisms are significant, insofar as they trigger

[4] *The Brains Blog* hosted a roundtable on "What can we learn from the Implicit Association Test?". Several of the authors in this volume contributed with very interesting comments (http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx; accessed April 27, 2017).

https://doi.org/10.1017/S2753906700002242 Published online by Cambridge University Press

clarifying questions that have the potential to enrich our understanding, for example, what does the IAT measure? What exactly do we learn from research on implicit bias? Does it help us predict people's behavior?

References

Amodio, David M., and Patricia G. Devine. 2006. Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. *Journal of Personality and Social Psychology* 91 (4): 652-61.

Anderson, E. 2010. *The imperative of integration*. Princeton: Princeton University Press.

Ayala, Saray. 2016. Agency and structural explanations of social injustice, presented at the Bias in Context conference, University of Sheffield, September.

------. 2017. Comments on Alex Madva's "A plea for anti-anti-individualism: How oversimple psychology misleads social policy." *The Brains Blog*. http://philosophyofbrains.com/wp-content/uploads/2017/03/Saray-Ayala-Lopez-Comments-on-Madva.pdf (accessed March 6, 2017).

Ayala, Saray, and Nadya Vasilyeva. 2015. Explaining speech injustice: Individualistic vs. structural explanation. In *Proceedings of the 37th annual conference of the cognitive science society*, ed. Rick Dale, Carolyn Jennings, Paul P. Maglio, Teenie Matlock, David C. Noelle, Anne Warlaumont, and Jeff Yoshimi. Austin, Texas: Cognitive Science Society.

Banaji, Mazharin, and Curtis Hardin. 1996. Automatic stereotyping. *Psychological Science* 7 (3): 136-41.

Brains Blog. 2017. Symposium on Alex Madva's "A plea for anti-anti-individualism." http://philosophyofbrains.com/2017/03/06/symposium-on-alex-madvas-a-plea-for-anti-anti-individualism.aspx (accessed March 10, 2017).

Devine, Patricia G., Ashby E. Plant, David M. Amodio, Eddie Harmon-Jones, and Stephanie L. 2002. The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology* 82 (5): 835-48.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the ethics of knowing*. Oxford: Oxford University Press.

Gigerenzer, Gerd, Peter M. Todd, and the ABC Research Group. 1999. *Simple heuristics that make us smart.* New York: Oxford University Press.

Gigerenzer, Gerd. 2010. Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science* 2 (3): 528-54.

https://doi.org/10.1017/S2753906700002242 Published online by Cambridge University Press

Haslanger, Sally. 2015. Social structure, narrative, and explanation. *Canadian Journal of Philosophy* 45 (1): 1-15.

------. 2017. Injustice within systems of coordination and cognition. *The Brains Blog*. http://philosophyofbrains.com/wp-content/uploads/2017/03/Sally-Haslanger-Comments-on-Madva-1.pdfhttp://philosophyofbrains.com/wp-content/uploads/2017/03/Sally-Haslanger-Comments-on-Madva-1.pdf (accessed March 6, 2017).

Lai, Calvin K., Patrick S. Forscher, Jordan Axt, Charles R. Ebersole, Michelle Herman, and Brian A. Nosek. 2017. A meta-analysis of change in implicit bias. *Open Science Framework*. https://osf.io/awz2p/ (accessed April 27, 2017).

Madva. Alex. 2016. A plea for anti-anti-individualism: How oversimple psychology misleads social policy. *Ergo* 3 (27): 701-28.

Saul, Jennifer. 2013. Scepticism and implicit bias. *Disputatio* 5 (37): 243-63.

------. 2017. Comments on Alex Madva's "A plea for anti-anti-individualism: How oversimple psychology misleads social policy." *The Brains Blog*. http://philosophyofbrains.com/wp-content/uploads/2017/03/Jennifer-Saul-Comments-on-Madva.pdf (accessed March 6, 2017).

Sherman, Benjamin. 2016. There's no (testimonial) justice: Why pursuit of a virtue is not the solution to epistemic injustice. *Social Epistemology* 30 (3): 229-50.

https://doi.org/10.1017/S2753906700002242 Published online by Cambridge University Press