**Review**

*Implicit Bias and Philosophy* (vol.1 & 2)

Edited by Michael Brownstein and Jennifer Saul

Oxford University Press, 2016, pp.640, *£*60

ISBN: 9780298713241

doi:10.1017/S00318191116000528

According to a surge of research findings it is all but undeniable that most people, including those who sincerely profess egalitarian beliefs, exhibit ‘implicit’ biases with respect to women, visible minorities, and other members of historically disadvantaged groups. These biases take the form of judgments or associations that, typically, lurk beyond introspective access and conscious control. Self-directed ‘stereotype threat’ is another major area of interest, in which thinking about one’s group identity might provoke anxiety, avoidance, and inhibit performance in stressful contexts, such as test-taking. This research rightly draws the attention of philosophers on a number of counts, no less because they might help explain ‘[p]ersistent inequalities between social groups’ (1: 1) and there are many facets to explore, including their nature, their role in decision-making, and the implications for self-knowledge, moral responsibility, and various aspects of social policy.

Some of these topics are covered in the first volume of this paired set, which advertises itself as a collection of papers bringing ‘a mixture of challenges to existing approaches and promising ways forward in the metaphysics and epistemology of implicit attitudes’ (1: 15). The second ‘integrates critical philosophical thinking with the latest empirical data’ in consideration of questions of responsibility and action (2: 7). Let us begin with an overview of what these works have to offer.

The first section of Volume 1 has interesting material on the character of implicit biases and how they figure into the structure of cognitive mechanisms. Some draw on influential ‘dual-process’ theories, such as Frankish (1.1) in consideration of self-control and akrasia, while Mallon (1.5) argues that although stereotype threat interferes with the idea of rational governance, everyday understandings of personhood are still important. Huebner (1.2) puts forward a three-part framework involving associative mechanisms and either model-based or model-free inferential processing. Turning to the nature of implicit biases, Holroyd and Sweetman (1.3) reject a simplistic division of associations into semantic and affective classes, concluding that further work is needed to determine the full variety of the phenomena. Machery (1.4) argues for a picture of implicit attitudes in which heterogeneous factors come together to produce something like character traits or dispositions, rather than mental states.

The second half examines questions about knowledge and rationality, including positions from Antony (2.1) contending that cognitive biases are an essential aspect of knowing and Madva (2.2) that we can avoid having to choose between being rational and being virtuous, though Goguen (2.3) takes a more pessimistic view about the costs of stereotype threat, such as self-doubt. The final three papers of Volume 1 deal with how implicit biases might contribute to disadvantage. For Hundleby (2.4) default assumptions about normalcy promote male privilege, for example by distorting perceptions of competence and fit for job applicants. Lee (2.5) takes a skeptical look at research challenging claims about sex discrimination in the sciences, while di Bella *et al*. (2.6) present data suggesting explicit associations between philosophy and maleness, though implicit links depended on the subject’s gender.

 Next, Volume 2 also contains useful discussion, this time on normative issues in three parts covering moral responsibility, structural injustice, and ethics. Washington and Kelly (1.1) contend that we can still be responsible for biases we are unaware of but ought to be on guard against. Glasgow (1.2) considers how to situate theories of action, especially about the Real Self, with respect to alienating biases, concluding that an alternative, content-sensitive ‘variantism’ taking account of non-agential features is preferable. In contrast, Faucher (1.5) discusses research supportive of Real Self views, but also suggests moderate revisions to our folk conceptions. Zheng’s piece (1.3) also pursues an agent-centered account, in effect arguing that biases not attributable to the Real Self can still call for accountability, along the lines of strict liability insurance practices. Sie and Voorst Vader-Bours (1.4) also argue we can be accountable for the effects of biases of which we are not directly responsible as individuals, though in terms of collective action. Extending this idea, the second section consists of a pair of essays taking individualistic and apolitical approaches to task. Blum (2.1) criticizes psychologists for downplaying the difference between stereotypes and true generalizations, as this hurts students and makes the need for systemic change less salient, and Jacobson (2.2) reflects on the complexity of injustice and need for change at multiple levels. The third and final section begins with a return to the individual, and the cultivation of bias-reducing strategies directed at internal factors, such as virtuous character traits (Rees 3.1), but also external ones, such as environmental context and situations (Browstein 3.2). Brennan (3.3) considers the accumulated effects of minor unintended harms and the concluding piece makes recommendations about British discrimination law (Hosking and Russell 3.4).

Among this impressive breadth in topics are thoughtful proposals about when we are responsible for implicit bias, what kind of responsibility is at issue, whether change should focus on individuals or institutions, and what form of action is called for. However, there are problems with both volumes concerning the handling of the empirical literature, which is sometimes made to jump to desired outcomes.

In a recurring error, hereby dubbed the ‘Glass Box' fallacy, logical leaps are made from crude behavioral measures to claims about specific cognitive information-processing mechanisms. In other words, black boxes are often treated as if their inner workings were transparent. A textbook-worthy example is a resume study that didn’t examine cognitive processes, but is treated by philosophers (in these volumes and elsewhere) as compelling evidence of implicit racism in hiring.[[1]](#footnote-1) This ignores such possible confounds as explicit racism and discrimination on the basis of socioeconomic status.[[2]](#footnote-2)

Another questionable practice is that presumptions about real-world impact are formed with little regard to the possibly artificial character of laboratory effects. Most of the authors also make normative recommendations of some kind based on the assumption that bias is bad. But if we aren’t sure what it is, or how it works, what exactly are we trying to overcome, and why? Instead of lumping all biases together as pernicious, Blum (2: 152ff.) emphasizes that there are ‘valid generalizations’ while Antony (1: 182) argues that the illusion of bias-free reasoning is a recipe for epistemic failure. So what is the percentage of implicit associations encoding true semantic content, such as that salt is often paired with pepper, or that men tend to be taller than women (2: 83 n. 33)? What is the percentage ofstereotypes that are accurate[[3]](#footnote-3)—do we have an obligation to disbelieve things that are true, but harmful? And how do we distinguish unjust, but very subtle, inequities—micro, nano and pico—from ‘mere inequalities’ (Brennan 2: 239)?

These, mostly neglected, difficulties are exacerbated by weaknesses in the science journalism, which a paragraph about anonymous review exemplifies (2: 195). A hiring audit about symphony orchestras is misinterpreted as establishing a causal relationship between ‘blind’ auditions and greater numbers of female musicians hired.[[4]](#footnote-4) Two further studies are presented as evidence of gender bias in journal refereeing, and though the first, Peters and Ceci (1982) is a seminal work,[[5]](#footnote-5) its context was institutional status, not gender. The second, Budden *et al.* (2008)[[6]](#footnote-6) was briefly hailed by *Nature* as ‘one bright light’ though a retraction soon followed with the editors admitting the paper ‘can no longer be said to offer compelling evidence of a role for gender bias’ adding they ‘cannot find other strong studies that support this claim’.[[7]](#footnote-7) As Peters and Ceci might have written, it is Frank Lee Manure.[[8]](#footnote-8) Oftentimes results are overstated or presented selectively, as when two well-known articles are claimed to jointly provide strong evidence of racial and gender bias in the evaluation of resumes (2: 62). It should have been noted that one of them did not find female names were downgraded, however. While some small errors are inevitable (and forgivable) in any scholarly work, the reviewer grew wary when confronted with an unfamiliar study.

Another case in point, raised over several chapters, is ‘shooter bias’ which ‘leads to disastrous consequences mainly for Black people’ (Sei and Voorst Vader-Bours, 2: 100) as in well-publicized police shootings (2: 24, n. 14; 2: 100, n. 18). Certainly some studies have produced evidence that race increases the proportion of ‘false alarms’—misidentifying unarmed targets as armed—perhaps starting with Correll *et al*.’s (2002, 1328)[[9]](#footnote-9) which was motivated by the worry that police ‘decisions to shoot’ interacted with ethnicity. But philosophers overlook the fact that most of these studies were conducted with undergraduates, not trained law-enforcement personnel. Indeed, while Brennan reports that Correll *et al.* (2007)[[10]](#footnote-10) found ‘White police officers and undergraduate students mistakenly shoot unarmed Black suspects more than White suspects’ in simulations (2: 244), the researchers noted ‘critical differences’ between the responses of police officers and other subjects (1017) in that the police ‘showed no bias’ in their decisions to shoot Black targets and were ‘dramatically’ less ‘trigger-happy’ than the others (1017; 1020). Brennan and sensationalistic news media confuse evidence for correlations between ethnicity and response timeswith *false positives*. For example, the news coverage of a meta-analysis (aggregating the results of 42 shooter-bias studies) conducted by Mekawi *et al.* reported their ‘corroborating’ the claim that shooters were ‘more likely’ to fire on Black targets[[11]](#footnote-11) despite the authors conceding on the key issue of false positives that subjects ‘were not, however, more likely to be sensitive to or have a higher false alarm rate for Black (versus White) targets’ (Mekawi *et al.* 2015, 128).[[12]](#footnote-12)

The difference it makes is on two counts: first, when it comes to understanding how implicit biases figure into cognitive architectures, and, second, when it comes to addressing very serious social ills. Concerning the former, research on racial stereotypes might be taken to suggest they enable slightly more rapid information processing (the disparity is one or two dozen milliseconds) if the target is stereotypical, with slightly slower processing if stereo-atypical, while having no significant effect on the rate of false alarms. This aspect of our associative mechanisms might well be a feature rather than a bug, e.g. on the assumption that the costs of the slower judgments are offset by gains in the faster judgments. Research is ongoing, though a recent study using ‘strikingly realistic’ simulators found that, notwithstanding their implicit biases, police officers were much less likely to shoot unarmed Black targets.[[13]](#footnote-13) The second issue concerns the explanatory value of shooter-bias for understanding why Blacks are so much more likely to get killed by the police. Writing in the *New York Times*, one expert points out that the surplus (more than two and a half times the baseline) is actually about where expected givenarrest rates.[[14]](#footnote-14) Adding to this, a recent working paper examined over 1000 police shootings across the United States and found that although police were more likely to use non-lethal means of force against Blacks, there was no evidence of racial bias in officer-involved shootings. The “only statistically significant differences by race demonstrate that black officers are more likely to shoot unarmed whites…” (Fryer, 34).[[15]](#footnote-15)

In light of these points, eliminating shooter-bias is unlikely to have much impact on the ghastly ‘killing gap’ that haunts the Black community, and solutions will depend on addressing massive social problems tied together by poverty, subpar education and housing, disparities in political power, drug use, crime, and easy access to firearms. While some authors make gestures along these lines, the absence of the perspective on shooter bias just sketched is not explained by the recentness of a work or two mentioned above. They tend to underappreciate when other factors might loom so large as to render implicit bias basically irrelevant (2: 248-9; 2: 176). An opportunity is missed to buttress calls for more action on structural injustices through careful consideration of research results.

One exception is Blum’s piece (e.g. 162), which also comes as a welcome relief from forays into scientifically dubious ‘micro aggressions’ (2: 75), bourgeois pearl clutching (2: 225), and toadying to kneejerk blame-and-shame mentalities (2: 37ff.). Likewise, for Jacobson’s (2: 2.2) vetting of Dixon’s skepticism about the efficacy of bias-reducing measures regarding persisting inequities, though in both cases greater attention to empirical detail would have made these discussions (as well as Brownstein’s consideration of ‘world-first’ strategies (2: 222-3) more rewarding.

Meanwhile, psychologists have been debating calls for closer scrutiny and higher standards, partly in light of replication failures, with some going so far as to speak of a ‘crisis’.[[16]](#footnote-16) Yet there is scarcely a hint of these concerns, which have been increasingly publicized since at least 2012.[[17]](#footnote-17) In almost two dozen essays, just shy of 600 pages of commentary, the reviewer noticed a single oblique reference to ‘controversial or uncertain aspects of the ongoing research’ (2: 17) and a footnote by Mallon (134, n.2). Why no consideration of the intense disagreement over the Stereotype Threat literature,[[18]](#footnote-18) Implicit Association Tasks,[[19]](#footnote-19) ‘P-fishing’,[[20]](#footnote-20) ‘P-Hacking’,[[21]](#footnote-21) the ‘File Drawer Problem’,[[22]](#footnote-22) and worries about reproducibility?[[23]](#footnote-23)

Arguably, philosophy is in the grip of its own kind of replication crisis in the sense of a tendency to take the empirical literature at face value. Within the narrowed spectrum there is room for some second thoughts—such as whether agent confusion might explain purported dissociations in explicit and implicit attitudes (2: 133), or that researchers conflate stereotypes concerned with performance with those concerning ability (2: 154), and in raising ambiguities about the interpretation of correlational studies—though only if they question claims about sex discrimination (2: 2.5). One can’t help but suspect many philosophers are impatient with the task of engaging inconvenient limitations of the research so they may hurry along to epistemic arguments for political ends, such as affirmative action initiatives (1: 1; 1: 168, 171-2; 2: 37-8; 2: 104; 2: 151; 2: 180; 2: 222). Despite these complaints the, sometimes quirky, topics raised—about coffee shop wait times, or whether white people smell like dogs when wet, or how lighting might affect racism—are ripe for further investigation. Actually, there is a second crisis, which is the failure to recognize the first, and this collection is at least illustrative of these trends.

1. E.g. (1: 2); (2: 18). [↑](#footnote-ref-1)
2. An attempted replication by Darolia *et al.* using surnames failed to find any significant differences in callbacks across race and gender, though once again what went through the minds of prospective employers is unknown. See Darolia, R., Koedel, C., Martorell, P., Wilson, K., and Perez-Arce, F. (2016). “Race and gender effects on employer interest in job applicants: new evidence from a resume field experiment,” *Applied Economics Letters* 23(12): 853-56. [↑](#footnote-ref-2)
3. Jussim, L. (2015). Précis of social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy. *Behavioral and Brain Sciences*, 1-66. [↑](#footnote-ref-3)
4. Despite Goldin and Rouse’s (2000) care in noting the limitations of their study, philosophers are united in committing this *post hoc* fallacy (e.g. 1: 166; 1: 248). See their *“*Orchestrating impartiality: The impact of ‘blind’ auditions on female musicians,” *The American Economic Review*, 90(4): 715-41. [↑](#footnote-ref-4)
5. Peters, D. P. and Ceci, S. J. (1982). “Peer-review practices of psychological journals: The fate of published articles, submitted again,” *Behavioral and Brain Sciences* 5(02): 187-95. [↑](#footnote-ref-5)
6. Budden, A. E., Tregenza, T., Aarssen, L. W., Koricheva, J., Leimu, R., and Lortie, C. J. (2008). “Double-blind review favours increased representation of female authors,” *Trends in ecology & evolution* 23(1), 4-6. This is also cited uncritically by Lee (1: 268). [↑](#footnote-ref-6)
7. M. Clarke, (2008) “No Demonstrated Gender Bias in Double-Blind Peer Review” *Peer to Peer*(blog) *Nature*, June 5. <http://blogs.nature.com/peer-to-peer/2008/06/no_demonstrated_gender_bias_in.html> [↑](#footnote-ref-7)
8. For an amusing account of their famous study, and the backlash against the researchers, see: https://thewinnower.com/discussions/7-the-peters-ceci-study-of-journal-publications [↑](#footnote-ref-8)
9. Correll, J., Park, B., Judd, C. M., and Wittenbrink, B. (2002). “The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals,” *Journal of personality and social psychology* 83(6): 1314-29. [↑](#footnote-ref-9)
10. Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., and Keesee, T. (2007). “Across the thin blue line: Police officers and racial bias in decisions to shoot,”

*Journal of Personality and Social Psychology* 92: 1006-23. [↑](#footnote-ref-10)
11. http://www.npr.org/2015/08/29/435833251/shooters-quicker-to-pull-trigger-when-target-is-black-study-finds [↑](#footnote-ref-11)
12. Mekawi, Y., and Bresin, K. (2015). “Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis,” *Journal of Experimental Social Psychology* 61: 120-30. [↑](#footnote-ref-12)
13. James, L., James, S. M., and Vila, B. J. (2016). “The Reverse Racism Effect,” *Criminology & Public Policy*. 15 (2): 457-79. [↑](#footnote-ref-13)
14. Mullainathan, S (2015). “Police killings of Blacks: Here is what the data have to say,” *New York Times* (Oct. 16). [↑](#footnote-ref-14)
15. Fryer Jr, R. G. (2016). “An empirical analysis of racial differences in police use of force” (No. w22399). National Bureau of Economic Research. [↑](#footnote-ref-15)
16. <https://osf.io/ezcuj/wiki/home/>; <https://www.psychologytoday.com/blog/the-nature-nurture-nietzsche-blog/201509/quick-guide-the-replication-crisis-in-psychology> [↑](#footnote-ref-16)
17. See <http://www.nature.com/news/nobel-laureate-challenges-psychologists-to-clean-up-their-act-1.11535> [↑](#footnote-ref-17)
18. <https://psmag.com/women-math-and-the-addition-of-stereotypes-e4fc1ea543cd#.s0qtebwg2>; <https://www.psychologytoday.com/blog/rabble-rouser/201512/is-stereotype-threat-overcooked-overstated-and-oversold> [↑](#footnote-ref-18)
19. <http://www.apa.org/monitor/2008/07-08/psychometric.aspx>; https://www.psychologytoday.com/blog/beautiful-minds/201101/does-the-implicit-association-test-iat-really-measure-racial-prejudice [↑](#footnote-ref-19)
20. <http://www.nature.com/news/scientific-method-statistical-errors-1.14700> [↑](#footnote-ref-20)
21. Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). “The extent and consequences of p-hacking in science,” *PLoS Biol*, *13*(3), e1002106. [↑](#footnote-ref-21)
22. Franco, A., Malhotra, N., and Simonovits, G. (2014). “Publication bias in the social sciences: Unlocking the file drawer,” *Science*, 345(6203): 1502-5. [↑](#footnote-ref-22)
23. Pashler, H., & Wagenmakers, E. J. (2012). Editors’ introduction to the special section on replicability in psychological science a crisis of confidence? *Perspectives on Psychological Science*, 7(6): 528-30. [↑](#footnote-ref-23)