Implicit Bias, Epistemic Injustice, and the Epistemology of Ignorance

Gus Skorburg

ABSTRACT: This paper uses the resurgence of interest in implicit bias research to revisit two significant developments in recent feminist epistemology: epistemic injustice and the epistemology of ignorance. I begin by arguing that Miranda Fricker's account of the virtue of testimonial justice finds empirical support in the research of social psychologist Margo Monteith. Next, I argue that while some findings in this body of research do license optimism about the prospects of correcting for the implicit biases at the heart of epistemic injustices, such findings need to be contextualized and situated in light of the insights generated in the epistemology of ignorance.

KEYWORDS: Feminist Epistemology, Epistemic Injustice, Epistemology of Ignorance, Implicit Bias, Social Psychology

WORD COUNT: 2,985

Introduction

Much has been written lately - in both academic and popular venues - about the operation of implicit bias, especially in the sphere of law enforcement. [1] The importance of this work can hardly be overstated, as the consequences and implications are far-reaching and immediately pressing. The rising tide of interest in implicit bias research provides numerous opportunities for philosophical reflection, and it is my purpose here to use this momentum to revisit two of the most significant developments in recent feminist epistemology: epistemic injustice and the epistemology of ignorance. Broadly, my aim here is to demonstrate the fruitfulness of a more robust dialogue between feminist epistemology and social psychological research on implicit bias. Following Miranda Fricker, I begin by arguing that the virtue of testimonial justice finds empirical support in the research of social psychologist Margo Monteith and her colleagues. I will not merely argue that Fricker's case is made out by appeal to empirical work, however. In the balance of the paper, I bring a feminist-inspired criticism to bear on this research. I argue that

[1] In the social psychological literature, for example, see Eberhardt et al. (2004) and Correll et al. (2007). For a popular take, see Chris Mooney's piece: "The Science of why Cops Shoot Young Black Men" <http://www.motherjones.com/politics/2014/11/science-of-racism-prejudice >

while some findings in social psychological research do license optimism about the prospects of correcting for implicit biases, such findings need to be contextualized and situated in light of the important insights generated in the epistemology of ignorance. Thus, I argue for a two-fold conclusion: feminist epistemologists can find useful resources in contemporary social psychological research, but insofar as we are genuinely interested in attempting to identify and eradicate the harmful effects of implicit bias, this research must be buttressed and kept in check by work in the epistemology of ignorance.

1. Epistemic Injustice and Implicit Bias

The central claim of Miranda Fricker's (2007) *Epistemic Injustice* is that "there is a distinctively epistemic kind of injustice…a wrong done to someone specifically in their capacity as a knower" (p. 1). Though she originally defines two such injustices - testimonial and hermeneutical - and many others have since been identified,[2] I will here focus on the testimonial variety of epistemic injustice. The link between Fricker's account of testimonial injustice and the social psychological research on implicit bias is the notion of a credibility judgment. Without rehearsing the details of Fricker's argument in detail here, suffice it to say that her account of testimonial injustice requires a commitment to the claim that, in testimonial exchanges, hearers usually make automatic, non-reflective, non-inferential judgments about the credibility of the speaker and the knowledge she is conveying. These non-inferential credibility judgments take the form of heuristics, stereotypes, and prejudices.[3] When the heuristics used in credibility judgments work well, the hearer accurately tracks relevant features of the speaker which signal their credibility or lack thereof. There are (at least) two ways in which these judgments can go

[2] For an overview of different varieties of epistemic injustice, see Hookway (2010).
[3] I hasten to add, as Fricker does, that these operations do not automatically or intrinsically carry a negative normative valence. Rather, as decades of cognitive scientific research indicate, the use of heuristics seems to be a fact about the way humans think. For an accessible introduction to this body of research, see Kahneman (2011).

wrong, however: the hearer can inflate the credibility of speaker, giving them more epistemic credit than is due or the hearer can deflate the credibility of the speaker, giving them less epistemic credit than is due. Fricker dubs the former credibility excess and the latter credibility deficit.

With the notions of credibility excess and deficit in place, we can make a pass at the paradigmatic case of testimonial injustice, which Fricker describes as follows: "the speaker sustains such a testimonial injustice if and only if she receives a credibility deficit owing to identity prejudice in the hearer; so the central case of testimonial injustice is *identity-prejudicial credibility deficit*" (p. 28).[4] With this brief sketch of the nature of testimonial injustice in place, we can now examine its relationship to social psychological research on implicit bias.

Testimonial injustice often happens at what Fricker calls the "spontaneous, unreflective level" (p. 89). This is precisely what makes testimonial injustice so insidious: at the level of reflection, conscious awareness, and beliefs, many people do not harbor prejudicial attitudes and beliefs (though some of course do). As Fricker correctly points out, these same people often do harbor "stealthier, residual prejudices, whose content may even be flatly inconsistent with the beliefs actually held by the subject. Certainly we may sometimes perpetrate testimonial injustice because of our beliefs; but the more philosophically intriguing prospect is that we may very frequently do it in spite of them" (p. 36). This is not a novel or controversial insight. Since the publication of Anthony Greenwald et al.'s seminal (1998) study using the Implicit Association Test (IAT), thousands of social psychological studies have uncovered discrepancies between

[4] Though much has been written about the various harms of testimonial injustice thus construed, for present purposes, suffice it to say that the primary harm of testimonial injustice is that "the subject is wronged in her capacity as a knower" (p. 44). That is, the speaker's standing in the community as capable of providing knowledge is undermined. An inability to provide knowledge, in turn, may imply an impoverished capacity for reason - an idea "played out by the history of philosophy in many variations, that our rationality is what lends humanity its distinctive value" (ibid).

subjects' explicit attitudes (e.g. "No group is naturally inferior to another group") and implicit associations (e.g. "white" with "good" or "black" with "bad") in domains ranging from skin tone, to weight, to gender, and to sexual orientation.

As Han et al. (2010) note, "the IAT has become the preferred implicit measure of for many psychological variables because implicit measures are presumed to be relatively immune from many of the concerns that plague self-report measures" (p. 2). In other words, the IAT is designed to detect precisely the kind of discrepancies that exist between reflective, explicit attitudes, and the stealthy, residual associations Fricker argues are at the heart of testimonial injustice. Insofar as testimonial injustice can thus be understood as involving the operation and expression of implicit biases, then the massive body of social psychological research on implicit bias ought to be of use to those interested in the phenomenon of epistemic injustice.

2. The Virtuous Hearer

Epistemic injustice occurs where credibility judgments break down, and hearers deflate the credibility of the speaker on the basis of identity prejudice. When this happens, the hearer is construed as failing to correct for the operation of credibility-irrelevant identity prejudice. In virtue theoretic terms, one might say that the hearer in such judgments is vicious. Faced with the same judgment, the hearer would be virtuous insofar as she "*neutralizes the impact of prejudice in her credibility judgments*" (p. 92). This neutralizing or corrective capacity is at the heart of the virtue of testimonial justice. I quote Fricker at length:

> What is needed on the part of the hearer in order to avert a testimonial injustice - and in order to serve his own epistemic interest in the truth - is a corrective anti-prejudicial virtue that is distinctively *reflexive* in structure. Such reflexive critical awareness of the likely presence of prejudice, then, is a prerequisite in the business of correcting for prejudice in one's credibility judgment. But what exactly is meant by 'correcting for' here? When the hearer suspects prejudice in her credibility judgments - whether through sensing cognitive dissonance between her perception, beliefs, and emotional responses, or whether through self-conscious reflection - she should shift intellectual gear out of spontaneous, unreflective move and into active critical reflection in order to identify how far the suspected prejudice has influenced her judgment (p. 91).

Importantly, what Fricker is *not* prescribing here is that we dispense with the heuristics and stereotypes that operate at the "spontaneous, unreflective level." It would be nearly impossible to do so. Rather, the project is to develop and sharpen the existing testimonial sensibility we already have in such a way as to minimize the effects of identity prejudicial credibility judgments. In much the same way the traditional Aristotlean ethical virtues are developed through habituation, so too are the epistemic virtues cultivated and reinforced through habits. The fully virtuous moral agent does not have to consult rules regarding the right way to construe a situation, the right emotions to feel, or the right course of action to take, but rather, they *just* construe, feel, and act in the way the situation requires. *Mutatis mutandis* for the virtuous hearer:

> The fully virtuous hearer, then, as regards the virtue of testimonial justice, is someone whose testimonial sensibility has been suitably reconditioned by sufficient corrective experience so that it now reliably issues in ready-corrected judgments of credibility. She is someone whose pattern of *spontaneous* credibility judgment has changed in light of past anti-prejudicial corrections and retains an ongoing responsiveness to that sort of experience. Full possession of the virtue, then, in a climate that has a range of prejudices in the social atmosphere, requires the hearer to have internalized the reflexive requirements of judging credibility in that climate, so that the requisite social reflexivity of her stance as hearer has become second nature (p. 97).

When this practice is sufficiently habituated, Fricker claims, the judgments made at the spontaneous, unreflective level will no longer need to be reflexively, consciously corrected for, but will simply be issued in a manner that has already corrected for the kinds of identity prejudices that cause testimonial injustices. But do we have any good reason to suspect that we are *actually* capable of doing this?[5]

3. The Social Psychological Findings

Fricker (2010) points to the work of Margo Monteith and her colleagues as providing empirical support for the virtue of testimonial justice. Monteith et al. (2002) propose a model for

[5] Here, I am following Alcoff (2010). Regarding the virtue of testimonial justice, Alcoff writes: "these are all volitional practices, or ones we might *consciously* cultivate and practice…if identity prejudice operates via a collective imaginary, as she [Fricker] suggests, through associated images and relatively unconscious connotations, can a successful antidote operate entirely as a conscious practice? Will volitional reflexivity, in other words, be sufficient to counteract a non-volitional prejudice?" (p. 132).

"putting the brakes on prejudice." Here, in rough sketch is their model: (1) Automatic stereotypes are activated and used. This leads to a (2) discrepant response in the agent. The agent (3) becomes aware of the discrepant response which leads to some combination of (a) behavioral inhibition, (b) negative self-directed affect, and (c) retrospective reflection. This (4) establishes *cues for control*. In the future, (5) the cues for control are activated when a possible discrepant response is perceived, and this in turn, leads to some combination of (a) behavioral inhibition and (b) prospective reflection which (6) inhibits the prejudicial response and generates an alternative response.

To illustrate this model in concrete terms, we are asked to imagine a white shopper who sees a Black person down the aisle. The white shopper (1) automatically assumes the Black person works at the store. When the shopper approaches and asks where to find an item, only to find out the Black person is also a fellow shopper, the white shopper (2) stops for a moment and (2b) feels embarrassed about the mistaken assumption. The white shopper (2c) makes note of how they felt and thereby links (2) the discrepant response with the (3b) negative affective response. This results in a kind of marker: (4) "when I'm in a similar situation in the future, I'll remember how embarrassed I felt and try to avoid the behavior that made me feel that way." So the next time the white shopper is in a store, seeing a Black person (or perhaps even seeing the original item they were searching for) will act as (5) a reminder that "this is the kind of situation that generated an undesirable response." Insofar as this is kept in mind, (6) the white shopper should then be motivated to avoid those kinds of undesirable thoughts and behaviors.

Monteith et al. (2002) ran a series of experiments to test the validity and tractability of this model. I will describe one experiment in detail here, though other experiments in the study also

provide similar support.[6] One series of experiments were designed to actually establish cues for control, and then test their efficacy — whether or not in situations for which cues for control had been established, inhibitory processes would be triggered. The task used for establishing the cues was the Implicit Association Test (IAT).[7]

In these experiments, participants took a variety of IATs and were presented with their results. Higher IAT scores indicate higher levels of prejudice (e.g. associating Black with unpleasant), and participants with greater IAT biases reported greater negative self-directed affects. Their hypothesis is that once participants became aware of their prejudices via their IAT score, this awareness would act as a cue for control in future tasks. Indeed, they found that the more participants experienced negative self-directed affects in response to their IAT scores, the more they paused in relation to Black names on subsequent IATs. "This suggests that the Black names were serving as cues for control" (p. 1044). And perhaps most importantly, "participants who felt guilty about their IAT performance were also less likely to generate racially biased responses. Because the participants had established cues for control, prospective reflection could occur, and racial biases were less likely to be manifested" (p. 1046).

It is not my intent here to draw general conclusions about the empirical tractability of Fricker's virtue of testimonial justice based on the results of one study. Rather, I only mean to

---

[6] For example, in the first series of experiments, the researchers manipulated participants' perceptions of whether or not they engaged in prejudiced responses via false physiological feedback procedures. In the experimental condition, participants received feedback that they were unable to control negative physiological responses to pictures of Blacks, leading them to believe they had prejudicial responses to the pictures. If it could be shown that some kind of behavioral inhibition was triggered in the experimental condition, that would provide a strong reason to suspect that explicit, volitional processes can access the prejudicial, implicit processes, given that the latter, as a result of the feedback manipulation, are so "stealthy" as to not even belong to the subject in the first place. And indeed, Monteith et al. found that "low prejudice individuals showed evidence of behavioral inhibition following prejudiced responses. That is, participants showed a brief interruption in ongoing behavior when presented with feedback that they had negative reactions to pictures of Blacks" (p. 1045).

[7] Monteith et al. describe it as follows: "This is a dual categorization task that assesses the extent to which automatically activated evaluations in relation to Blacks are less pleasant than in relation to whites. More specifically, the task measures the strength of association between pleasant and unpleasant words (e.g. *sunshine* vs. *stink*) and, historically, white versus Black names (e.g. *Adam* vs. *Tyrone*)" (p. 1042).

show that insofar as testimonial injustice can be understood as bound up with the operation of implicit bias, the outcomes of interventions developed by social psychologists to combat implicit bias do give us a reason to suspect that similar techniques may also be effective in combating epistemic injustice. And indeed, other studies, such as Devine et al. (2012), seem to suggest that we can be much more sanguine than this.[8] In the last section, I will argue, however, that there are also at least as many reasons for skepticism about the efficacy of such interventions.

4. Challenges from the Epistemology of Ignorance

Charles Mills' (1999) *Racial Contract* is widely recognized as the first formulation of the idea of an epistemology of ignorance.[9] In a lesser known passage, but one very much of interest to the present paper, Mills writes: "what we need to do, then, is to identify and learn to understand the workings of a racialized ethic. How were people able to consistently able to do the wrong thing while thinking they were doing the right thing? In part, it is a problem of cognition and of white moral cognitive dysfunction. As such, it can potentially be studied by the new research program of cognitive science" (pp. 94-5).

While we may no longer think of cognitive science as a new research program, Mills' insight still rings true. Indeed, Allison Bailey is on exactly the right track when, referencing the same passage, she writes: "Mills does not give readers much detail here and, to be fair, this is not his project. However, the Harvard Implicit Association Test offers an example of what I think he

[8] By their own lights: "our own results provide compelling and encouraging evidence for the effectiveness of our multifaceted intervention in promoting enduring reductions in implicit bias. As such, this study provides a resounding response to the clarion call for methods to reduce implicit bias and thereby reduce the pernicious unintended discrimination that arises from implicit biases. Reductions in implicit bias that emerged by week 4 following the interventions are unprecedented in the literature" (Devine et al. 2012, p. 1276).

[9] The most famous formulation reads: "the racial contract prescribes for its signatories an inverted epistemology, an epistemology of ignorance, a particular pattern of localized and global cognitive dysfunctions (which are psychologically and socially functional), producing the ironic outcome that whites will in general be unable to understand the world they themselves have made" (Mills 1999, p. 18). For a contemporary overview of the epistemologies of ignorance, see Alcoff (2007).

has in mind" (p. 119). After briefly describing how the categorization task works, Bailey continues,

> quick responses to these pairings reveal subjects' implicit attitudes. From there it is a short step to asking how these preferences influence moral deliberation. If associations such as white = glorious are learned, then they can be unlearned. Perhaps this is what Mills has in mind when he encourages people to think against the grain and to 'learn to trust [our] own cognitive powers, to develop [our] own concepts, insights, modes of explanation, overarching theories, and to oppose the epistemic hegemony of conceptual frameworks designed in part to thwart and suppress the exploration of such matters' (Bailey 2007, p. 81, quoting Mills 1997, p. 119).

All of this might seem to suggest that Fricker's appeal to the work of Monteith and her colleagues would be very much in line with the projects of the epistemology of ignorance as set out by Mills. Bailey, however, also expresses what I take to be a very pressing concern:

> Whites wanting to undo our ignorance can work hard to reverse the biases revealed to them by the Implicit Association Test. We can thumb through volumes of history to reveal stories that have been kept from us. We can engage in both of these activities from the safety of our worlds. These solutions offer a temporary remedy to white cognitive dysfunction, but they do so in ways that rely on isolated, noninteractive, self-reflective, and solipsistic processes" (p. 90).

I would also add that the intervention outcomes cited by Monteith and others are most significant for already low-prejudice college students - those who are often motivated to be unbiased in the first place. We must also ask how effective are the cues for control for individuals who are not highly motivated to avoid prejudicial judgments and behavior in the first place? And this is to say nothing of the limitations of experimental conditions: are there any good reasons to suspect that the cues for effective behavioral inhibition extend beyond the scope of the experiment? That is, are they effective in the long term? Are there any good reasons to think that the cues which control such behavioral inhibitions are generalizable to novel situations?

To sum up the foregoing worries, I want to suggest that the literature on the epistemology of ignorance forces us to take seriously the possibility that most people are not low-prejudice to begin with; they are perhaps not motivated to address their prejudices, even if they are made

aware of them. Not only this, but significant social, economic, and political incentives accrue to those in positions of power and privilege who *actively ignore* their prejudices and biases. Many, if not most of us, are strongly motivated to *not* examine our prejudices, much less attempt to eradicate them.

Conclusion

On the one hand, we seem to have promising evidence that interventions developed by social psychologists actually do reduce implicit biases. On the other hand, we are presented with forceful arguments from the epistemology of ignorance that such interventions may be quite limited in their scope. If this paper has succeeded in its aims, then it will have raised more questions than it answers. Can the myriad operations of epistemic injustice be studied empirically? Is the active production of ignorance the kind of thing that psychology can study? If so, how? Moreover, are the methodologies of psychology or philosophy robust enough to prescribe solutions to these problems?

I do not believe that philosophers alone, nor psychologists alone can begin to answer these questions. I do believe that a philosophical account of bias uninformed by empirical research is empty, while empirical research on bias uninformed by critical philosophy is blind. Perhaps the only definitive conclusion that is supported by the arguments in this paper is that collaboration between philosophers and social psychologists, both working in a feminist vein, is the best place to begin addressing the pernicious phenomena of implicit bias, epistemic injustice, and ignorance.

Works Cited

Alcoff, L. (2010). Epistemic Identities. *Episteme*. 7(2), 128-137.

Alcoff, L. (2007). Epistemologies of Ignorance: Three Types. *Race and Epistemologies of Ignorance*, Sullivan and Tuana, eds. State University of New York Press. 39-58.

Bailey, A. (2007). Strategic Ignorance. *Race and Epistemologies of Ignorance*, Sullivan and Tuana, eds. State University of New York Press. 77-94.

Correll et al. The Influence of Stereotypes on Decisions to Shoot. *European Journal of Personality and Social Psychology*. 37(6), 1002-1117.

Devine et. al (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*. 48, 1267-1278.

Eberhardt et al. (2004). Seeing Black: Race, Crime, and Visual Processing. *Journal of Personality and Social Psychology*. 87(6), 876-893

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Fricker, M. (2010). Replies to Alcoff, Goldberg, and Hookway on *Epistemic Injustice*. *Episteme* 7(2), 164-178.

Goldberg, S. (2010). Comments on Miranda Fricker's *Epistemic Injustice*. *Episteme*. 7(2), 138-150.

Greenwald et al. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*. 74(6), 1464-1480.

Han et al. (2010). Malleability of attitudes or malleability of the IAT? *Journal of Experimental Social Psychology*. 46, 286-298.

Hookway, C. (2010). Some Varieties of Epistemic Injustice: Reflections on Fricker. *Episteme.* 7(2), 151-163.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Mills, C. (1999). *The Racial Contract*. Cornell University Press.

Monteith, M. et al. (2002). Putting the Brakes on Prejudice: On the Development and Operation of Cues for Control. *Journal of Personal and Social Psychology*. 83(5), 1029-1050.

Mooney, C. (2014). The Science of Why Cops Shoot Young Black Men. <http://www.motherjones.com/politics/2014/11/science-of-racism-prejudice >