Category Matters: The Interlocking Epistemic and Moral Costs of Implicit Bias

In this paper I reject the claim—made both by Tamar Szabo Gendler in On the Epistemic Costs of Implicit Bias *and Jennifer Saul in* Scepticism and Implicit Bias*—that in order to be epistemically and morally responsible, social categories should not influence our evaluations of individuals or subsequent actions. I will provide evidence against the claim by denying its empirical plausibility, emphasizing the epistemic and moral benefits that may come from social categories, and reconceptualizing the inclusion of base-rate information. Throughout the paper I will emphasize the unique interlocking of epistemic and moral considerations that are relevant to implicit bias, bias mitigation, and responsibility. It is my hope that this analysis lays the groundwork for an account of the right ways social categories can affect our judgments, i.e. the ways in which such influence may improve our epistemic and moral situations rather than degrade them.*

1. Introduction

On October 26, 2016 the Young Conservatives of Texas (YCT) at The University of Texas at Austin held an anti-affirmative action[1] bake sale with prices based on race and sex, charging more to individuals with social identities that YCT argued benefit the most from affirmative action. The Chairman of the organization, Vidal Castañeda, stated, "Our protest was designed to highlight the insanity of assigning our lives value based on our race and ethnicity, rather than our talents, work ethic, and intelligence…It is insane that institutional racism, such as affirmative action, continues to allow for universities to judge me by the color of my skin rather than my actions."[2] Put another way, the justification offered for the bake sale is that social category membership should be irrelevant to judgments and decisions about individuals.

This is a particular kind of argument, made for a particular political purpose, in hopes of a particular outcome. Notably, philosophers writing on implicit bias, philosophers with far different goals than the YCT, also utilize a version of this same claim: social categories *should* be irrelevant to our assessments of individuals. I will call this claim The Irrelevance Assumption.[3] Philosophers then add the following premises to make arguments about pernicious epistemic[4] mistakes:

> P1 (The Irrelevance Assumption): Social categories should be irrelevant to our assessments of individuals.

[1] It is not the purpose of this paper to take a stance on affirmative action. My point will be to show that the same problematic premise can be used as the foundation for many different types of arguments.
[2] For more information see www.cnn.com/2016/10/28/us/university-bake-sale-trnd/.
[3] The claim is formulated and used differently by various theorists, including the two philosophers I will focus on, Saul and Gendler. These differences do not, however, change the content I wish to target.
[4] To shift the focus of the argument to the moral costs of implicit bias, one need only to replace 'epistemic' with 'moral' in the argument. Both versions are common in the literature. My purpose here is to demonstrate a type of argument in which The Irrelevance Assumption may be used, rather than a particular view.

P2: When a person harbors implicit biases and those implicit biases influence an assessment, then a social category has influenced that assessment.

P3: Including irrelevant information in assessments is an epistemic mistake.

Conclusion: Implicit bias leads to epistemic mistakes.

In this paper I argue that The Irrelevance Assumption is mistaken, or that there are at least some cases in which social categories are relevant to our assessments and should be taken into account. Following Charles Mills (2000), who, in detailing the metaphysics of racial categories, claims, "That race *should* be irrelevant is certainly an attractive ideal, but when it has *not* been irrelevant, it is absurd to proceed as if it had been" (41)[5], I will argue that rather than aspiring to ignore, disregard, or overcome these influences, we must take social categories into account if we are to be in the best moral and epistemic positions. Though my aim here is primarily to give a negative account urging for the rejection of The Irrelevance Assumption, it is my hope that this paper clears the ground for a positive view about the right ways social categories might influence our judgments about a person or group of people. In other words, I want to show it is not *that* a social category influences our judgment that is important, but rather *how*.[6]

To analyze this assumption, I will focus on two papers: *On the Epistemic Costs of Implicit Bias*, written by Tamar Szabo Gendler (2011), and *Scepticism and Implicit Bias*, written by Jennifer Saul (2012). The motivation for these selections is twofold: one, both constitute novel and major contributions to the philosophical work on implicit bias, particularly regarding the special connection between the epistemic and the ethical in the identification and mitigation of implicit bias; two, the assumption that social categories should not influence our judgments of individuals is important for each argument, yet left undefended. In section 3, I clarify the work this assumption does for each argument in turn. I want to note that rejecting The Irrelevance Assumption does not entail rejecting the conclusion of either paper (though the conclusions will require different arguments). Rejection of The Irrelevance Assumption forces a reconceptualization of the guiding questions for research on the potential mitigation of implicit bias. It is not my hope to give a full account of moral responsibility for implicit bias.[7] Rather I will challenge a claim that has not been

[5] Though there are certainly differences between types of social categories, particularly with respect to historical legacy, many of his arguments can be applied to other categories such as gender or sexual orientation.

[6] Alex Madva (2016) puts forth a similar view and explores the possibility of "*regulating* the accessibility of our social knowledge in order to have that information available when and only when we need it" (201).

[7] Those interested in these accounts should see Brennan (2016), Brownstein (ms), Rees (2016), Fricker (2015), Glasgow (2016), Holroyd (2012), Levy (2012), Washington and Kelly (2016), and Zheng (2016).

defended, yet plays a key role in important, agenda-setting arguments, but which itself is, I argue, questionable.

2. Implicit Bias

In this section I will give a brief descriptive account of the contemporary literature on implicit bias. The term implicit bias refers to the unreflective and hard to introspectively access set of automatic associations that may lead to prejudiced judgment and behavior.[8] Interest in implicit bias was triggered by psychological findings about the nature of implicit association. The development of the Implicit Association Test (IAT; Greenwald, et al., 1998), an indirect measure of implicit associations between two target concepts, and subsequent findings pushed psychologists and philosophers to think about possible tensions between reports of belief and unconscious associations.

Over 75% of Americans who have taken the Race IAT show an automatic preference for White faces over Black faces (Banaji & Greenwald, 2013). In addition, studies that collect both implicit preferences (through the IAT) and explicit preferences (through self-report measures) show that White participants have greater implicit preferences for White over Black ($d = .83$) than explicit preferences for White over Black ($d = .59$) (Nosek, et al., 2002). This demonstrates both the strength of the preference, as well as the discordance between reports of belief and implicit associations. Since the development of the IAT, several additional indirect association measures have been developed. Some examples are the Affect Misattribution Procedure (AMP; Payne et al. 2005), the Go/No-go Association Task (Nosek & Banaji, 2001), and the Extrinsic Affective Simon Task (De Houwer, 2003).

Further, there is evidence that implicit bias is correlated with prejudiced behavior. Since 2007, many IAT studies have also included a behavioral measure to test the IAT's predictive validity (Banaji & Greenwald, 2013). A 2009 meta-analysis showed that IAT effectively predicts a range of prejudiced behavior ($r = .274$) and does so better than self-report measures for socially sensitive issues such as race (Greenwald, et al.). These studies addressed challenges claiming that the IAT may not be correlated with behavior or not be correlated more strongly than explicit report measures. Taken together these studies suggest something troubling: we likely have implicit associations that

[8] Though an interesting and worthwhile pursuit, the exact mental nature of implicit bias will not be explored in this paper. The arguments made in this paper will be relevant whether implicit biases are beliefs (Mandelbaum, 2016), aliefs (Gendler, 2008), *FTBA* attitudes (Brownstein, ms), or character traits (Machery, 2016).

affect our judgments and behavior in ways that we would explicitly disavow. In other words, individuals may have strong commitments to racial equity,[9] but think and act in ways that notably work against this goal.

3. Gendler and Saul

In this section I will sketch the arguments given in Gendler (2011) and Saul (2012), highlighting their uses of The Irrelevance Assumption: that social categories should not influence our judgments of individuals. For both Gendler and Saul, epistemic and moral costs are closely tied; however, the relationship between the costs is different for each. For Saul, as epistemic costs are mitigated, so are moral costs; the costs are directly related. For Gendler, on the other hand, as moral costs are mitigated, epistemic costs are incurred; the costs are inversely related.[10] This differing relationship influences the way each philosopher utilizes The Irrelevance Assumption. Though Gendler's *On the Epistemic Costs of Implicit Bias* came before, I will discuss Saul's *Scepticism and Implicit Bias* first as she explicitly states The Irrelevance Assumption.

In this insightful and influential article, Saul claims that contemporary research about implicit biases and their effects gives rise to a new kind of skepticism, a 'bias-related doubt'. Though this doubt doesn't lead us to question the existence of the external world or other minds like traditional forms of skepticism, we do have "very good reason to believe that we cannot properly trust our knowledge-seeking faculties," particularly when it comes to our knowledge about other people, their capacities, and intentions (243). She claims that this type of skepticism is in some sense stronger than traditional skepticism because it "demands action" (243). Doubting the existence of our hands doesn't prompt us to radically change our epistemic or moral situations; doubting the accuracy of our everyday credibility and like judgments does. She argues for this conclusion by giving a series of troubling empirical cases in which implicit bias plays a role—CV evaluation, journal submission evaluation, and shooter bias—and highlights the moral, political, and epistemic consequences of the impact of implicit bias. She then gives a variety of possible solutions for improving our epistemic and moral situations.

I agree with Saul. I think there is something going wrong morally and epistemically when our negative and inaccurate implicit biases affect our actions and judgments. However, I find one

[9] Other formulations of this commitment, such as to egalitarianism or treating people equally, will also pose a similar problem. The inconsistency between the avowal and the implicit association is the interesting phenomenon, not the particular nature or wording of the avowal/commitment.

[10] Particularly with respect to the encoding and use of relevant base-rates.

assumption she makes throughout the paper troubling. Here I'll give several instances of the assumption:

> "These judgments are very clearly being affected by something that *should* be irrelevant—the social category of the person…" (244).
>
> "…they shouldn't be looking at the credibility of an individual at all…we are likely to be affected by the social group of the person presenting evidence or an argument even when we were [sic] are trying to evaluate that evidence or the argument itself" (249).
>
> "These mistakes are ones in which something (the social category of the individual) that we actively think *should not* affect us does" (249).
>
> "if you actually are basing lots of decisions on the social categories that people you encounter belong to, then you're clearly not doing as well as you can" (256).

To ensure I am not being uncharitable, I want to clarify what I think she might mean: that we are making mistakes when our *inaccurate* stereotypes about social categories alter our judgments and actions. She points to this when she details some potential mistakes:

> You're making the wrong decisions epistemically speaking: taking an argument to be better than it is, perhaps; or wrongly discounting the view of someone you should be listening to. You're also making the wrong decisions practically speaking: assigning the wrong mark to an essay, or rejecting a paper that you should accept. Finally, you're making wrong decisions morally speaking: you are treating people unfairly; and you are basing your decisions on stereotypes that you find morally repugnant (256).

In these cases, unconscious social category biases influence judgments in a way that makes one less accurate and less likely to acquire desired knowledge. However, I want to emphasize that these mistakes are not *merely* a result of social categories influencing judgment or action, but rather of erroneous and pernicious social category associations altering judgments or actions. It is not *that* a social category affects judgment, but *how*. Though this final quotation leads us in this direction, the above four selections do not. In those, the assumption is more baldly stated, i.e. that no matter how the social category influences our judgment and decisions, we've made a mistake. It is this assumption, The Irrelevance Assumption, I want to reject.

I will now turn to Gendler's discussion of epistemic costs. Though similar in topic, the argument put forth leads to a vastly different conclusion. Rather than asserting that we must do

something to avoid the types of epistemic and moral mistakes that arise from implicit bias (and that this may indeed be possible with the right, but not-yet empirically discovered, kinds of strategies), Gendler concludes, "living in a society structured by race appears to make it impossible to be both rational and equitable" (57). That is, if we mitigate the moral costs of implicit bias, we increase the epistemic costs, and visa versa.

Throughout the paper, Gendler highlights epistemic costs associated with implicit bias, three that result from the phenomenon itself, as well as living in a racialized environment in general (this discussion is where she is most closely aligned with Saul), and one from attempts at mitigating the effects of implicit bias. She emphasizes throughout the discussion that these costs are incurred regardless of whether the individual avows the content of her automatic associations.

First, she identifies the cross-race recognition deficit in which individuals are more likely to remember specific facial features of own-race individuals than other-race individuals. Rather than encoding information that would allow future recognition, participants encode the face as "racial category" for the purposes of classification. She asserts that the tendency to encode this way is a result of automatic associations. Second, Gendler describes stereotype threat: "a well-documented phenomenon whereby activating an individual's thoughts about her membership in a group that is associated with impaired performance in a particular domain increases her tendency to perform in a stereotype-confirming manner" (48). Negative implicit biases turned inward lead to epistemic costs for particular tasks.[11] Third, she gives an account of cognitive depletion after interracial interaction; after white participants interacted with a different-race peer, they performed more poorly on executive control tasks. Cognitive depletion is particularly high on this task for those who have avowals that are discordant with their automatic racial preferences (implicit biases), indicating that the depletion may be caused by attempts to suppress automatic behaviors stemming from these biases. In describing these costs, Gendler's account is similar to Saul's: implicit biases lead to epistemic costs and further behavior that may not be in-line with avowed anti-discriminatory commitments.

Gendler then turns to epistemic costs associated with mitigating implicit bias; this is where her argument parts ways with Saul's. However, she still makes use of the claim I wish to reject, though she gives it a different role. In this discussion Gendler cites research on what Philip Tetlock, et al. (2000) call "forbidden base rates" to claim that mitigating implicit bias often requires one to

[11] When the biases are positive, performance may improve. This phenomenon referred to as Stereotype Lift. For further research, see Froehlich, et al. (2016).

ignore important social category information that may *improve* our epistemic situations, rather than degrade them. For example, citing Tetlock, she gives cases in which individuals did not take race-correlated actuarial risk into account when assigning insurance premiums and "engaged in a kind of epistemic self-censorship on non-epistemic grounds" (55). She categorizes this censorship behavior as irrational, even though it aligns with anti-racist avowals. It is here that epistemic and moral concerns are in tension with one another.

The Irrelevance Assumption shows up in her discussion of bias mitigation or what one *might do* to avoid epistemic and moral costs.[12] She asserts that to reduce epistemic costs we might "fail to encode the base rate information and cultural associations that give rise to these problematic aleifs [Gendler's term for implicit attitudes]" (54). Because I think it is empirically unlikely that we can "fail to encode" associations,[13] I'll rephrase: one might keep the social category of an individual from affecting judgments and subsequent decisions in order to improve epistemic and moral outcomes. You might think that this rephrasing is too self-serving; however, Gendler also suggests that we might ignore social category base-rates for ethical reasons (i.e. upholding anti-racist commitments), which seems like a clear case of The Irrelevance Assumption and related argument-type given in Section 1. Because she thinks that there are times in which relying on social categories improves our epistemic situation, Gendler's use of the assumption is slightly different than Saul's; nevertheless, both use The Irrelevance Assumption as the ideal for those wishing to improve epistemic and moral situations with respect to implicit bias.

4. Rejection of The Irrelevance Assumption

In this section of the paper I show that The Irrelevance Assumption is false, i.e. that there are circumstances when, in considering another person, information about the social categories that person belongs to is not only relevant to, but also *should be used* in the assessment of that person. My treatment of each challenge will be short, and it is my hope that these critiques provide the ground for a continued discussion. First, I will challenge the premise on empirical grounds, asserting that it is likely not possible to make evaluations independently of social category information. Second, I will demonstrate two epistemic benefits that arise from taking social categories into account. And, third, I will discuss base-rate neglect, emphasizing that the inclusion of negative base-rates (such as

[12] Thus, it may be unfair to attribute the assumption *to* Gendler. However, it seems like Gendler would've given another option for bias mitigation, if she thought one was available.
[13] This will be a part of my rejection of The Irrelevance Assumption. See section 4.1 for details.

crime statistics) is not the only plausible way to include social category information. An important upshot of this reframing will be that Gendler's conclusion is mistaken; it will be possible to be both rational and equitable.[14]

4.1 Empirical Possibility

Taking into account the psychological literature on implicit encoding and associative attitudes, we may wonder whether it is possible to keep social categories from influencing evaluations of individuals. I will take a familiar stance on responsibility here: it seems odd to require something that is not possible, even if—were it possible—it would be ideal. The empirical challenge can be mounted from two fronts: the automatic encoding of stereotype information and the automatic tendency to group individuals into social categories and apply category relevant information. More than half a century ago Gordon Allport (1954) described categorization as a basic feature of effective cognitive functioning. [15] Perhaps most telling is the early age at which individuals begin to use social categories to understand the world. Children as young as three or four already use gender and race in reasoning tasks (Shutts, et al., 2012). Particularly enlightening are discussions of human kinds and their development over time and space (Hacking, 1995; Mallon, 2016; Mills, 2000)[16], which suggest that most of our social interactions are structured by social category thinking. Further, much of the research on implicit bias itself supports the automatic encoding and use of social categories and that these categories update based on new information and experiences (Brownstein, ms). Further empirical and theoretical research on social category cognition and its mechanisms may allow us to shift the ways in which social category information is encoded and made salient for use in unconscious processing, explicit reasoning, and judgment. My point here is not that we shouldn't worry about these processes because they are automatic and unavoidable, but rather that we ought not make the mistake of assuming it is possible to ignore social category information.[17]

4.2 Epistemic Benefits

In this section, I will discuss two epistemic benefits that arise from including social category information, at least when it's done right: increased testimonial credibility[18] and robust social

[14] Joshua Mugg has taken on this assertion as well; see his 2013 paper and current manuscript for details.
[15] For a similar argument with a direct application to implicit bias, see Antony (2016).
[16] The references I list here are philosophers utilizing large swaths of empirical literature to make philosophic arguments, rather than original empirical research.
[17] Thanks to an anonymous reviewer for pushing me to clarify this point.
[18] This example is prompted by Miranda Fricker's 2007 account of Epistemic Injustice, i.e. harms done to individuals in their capacity as knowers. Though not discussed by Fricker, it is reasonable to assert that implicit biases of the type

exchange. First, the social category of a speaker should increase a hearer's credibility judgment of a speaker when the social category is relevant to the content of the testimony. One of the most common critiques of reproductive policy makers in America is that they are making arguments and decisions about something they will never experience; male representatives are making laws that determine the decisions women can make about their bodies and family planning.[19] Similarly, a common critique of policy makers and political leaders from activist groups like Black Lives Matter is that leaders make claims and decisions about black lives, even when they fail to understand the experience of black women and men.[20] In light of these critiques, one clear way social categories can play an important and valuable role in testimony evaluation is in relationship to the *content* of the testimony. Intuitively, it makes more sense to avow the testimony of someone giving evidence about common experiences of members of their *own* social category than the testimony of someone speaking about experiences had by those in other social categories. Further, we might expect that members of oppressed social categories have privileged insight and a more developed critical lens through which to see our social and political sphere, particularly with respect to social inequities they themselves experience (Collins, 1990; Harding, 1991; Mills, 1998, Smith, 1974).

Second, the influence of social category information on decisions such as hiring and academic admittance ensures a robust intellectual and work community that encourages vibrant discussion and multiple perspectives.[21] The ability of diverse groups to come to superior decisions because of deliberation[22](Landemore, 2012) means that, to improve the epistemic situations of groups, one ought to pay attention to the social categories of individuals that will learn or work together. Another benefit of this strategy is that it mitigates at least some worries about CV (Moss-Racusin, et al., 2012) and like evaluation so often cited in the philosophical literature, thereby lowering epistemic and moral costs associated with implicit bias. Admittedly, both examples are of explicit reasoning and decision-making about the inclusion of social category information, rather than of the unconscious influences of pernicious implicit biases about which Saul and Gender are

---

described above can be responsible for, or produce, the kind of social identity prejudice necessary to set credibility lower than it should (based on relevant factors such as expertise, experience, etc.) be set. This leads to a testimonial epistemic injustice. For more detail, see Fricker (2007).

[19] Though I don't have space to detail the history of this critique, here is an example of a recent protest from my home state, Indiana: http://www.nytimes.com/2016/04/08/us/periods-for-pence-campaign-targets-indiana-governor-over-abortion-law.html

[20] Similarly, I cannot give a full account. Here is an example: http://www.theroot.com/articles/culture/2014/08/ferguson_how_white_people_can_be_allies/

[21] For a more robust account of the epistemic value of diversity, see Robertson (2013).

[22] For a critique of this position, see Stich (2014).

worried. My point here is to give evidence against and reject The Irrelevance Assumption, which is agnostic about the reasoning process used to assess individuals.

4.3 Base-Rate Neglect

A further discussion of Gendler's base-rate neglect is in order. It may seem that above I simply agree with Gendler's conclusion: that social category information, such as base-rates, should be included in our judgments of others upon penalty of irrationality. A tacit assumption in these discussions of base-rates is that social category information only provides negative information about individuals and improves our epistemic situations by telling us whom to avoid or whom not to trust. In the previous two sections, I've given some evidence that challenges this claim; paying attention to social categories improves our epistemic situations by signaling expertise, ensuring a variety of perspectives, and limiting epistemic mistakes.

Further, I want to comment on the purported irrationality of ignoring accurate base-rate information for practical purposes (in this case anti-racist purposes), beginning with an example that Gendler uses to describe alief. To emphasize the tension between automatic "representational-affective-behavioral" aliefs and avowed commitments (Gendler calls them endorsed beliefs) she details the fear one may feel on a skywalk above the Grand Canyon. Although one desires to view the Grand Canyon in the suspended position and avows that the skywalk is safe, one may experience intense fear upon ascending into the clear case. If one is able to brave the Skywalk regardless, it is likely because one's avowals and desires (to view the Canyon) overcome the automatic reaction. I think this is similar to base-rate neglect. It is not clear that the Skywalk override is rational on Gendler's view. There *is* a chance, however slight, that the bridge will break.

If you don't think the Skywalk example demonstrates this, take riders of rollercoasters or carnival rides. Most have seen media stories of twelve passengers hanging upside down in a broken down rollercoaster or, if you'd like a more gut-retching example, someone's legs smashed when a mechanism fails. Typically, we do not say that these individuals have been irrational for ignoring their fears and continuing to experience the ride. Rather, we would say that they were irrational if their fears *kept* them from riding the rides. We could also push this example further to everyday activities that are, according to base-rates, very dangerous, but in which are seemingly not irrational to engage. Take driving. Individuals who drive on a daily basis are continually putting themselves at risk. On Gendler's picture of rationality and base-rate neglect, if we have encoded base-rates correctly, then it seems the decision to ignore the base-rate risk of driving is engaging in irrational behavior; we suffer an epistemic cost for a practical reason (i.e. driving is the most convenient mode

of transportation for most people). Couching base-rate neglect in objectively rational standards not only dangerously simplifies the multiple cares and concerns of individuals, but is also sensitive to these and like counterexamples. In this section I have analyzed Gendler's conception of base-rate neglect to show that conceptualizing social category information as providing only negative information about individuals and focusing on an objectively rational application leads us astray. Rethinking the ways social categories may influence our assessments of individuals gives us further reason to reject The Irrelevance Assumption and provides evidence against Gendler's conclusion that we must choose between being rational and moral.

5. Conclusion

Most troubling about the suggestion that we should render social categories irrelevant to our evaluations of others is that it seems to commit one to a sort of in-principle colorblindness[23] and lack of cultural awareness. It also bars one from taking a culturally nuanced and intersectional approach to addressing systems of social inequity. Although irrationality may be the ultimate sin for philosophers, it seems this other sort of worry is far more important for those committed to building an equitable world. In this paper I have provided evidence against The Irrevelavance Assumption, the claim that social categories are or should be irrelevant to our evaluations of individuals. Further, I have provided some positive reasons to demonstrate that social categories can play an epistemically and morally productive role. The rejection of The Irrelevance Assumption does not lead to a full-stop rejection of the conclusions of either paper; rather, it leads us to reframe questions about implicit bias mitigation, as well as positive ways to move forward until empirical methods are developed. When we conceptualize, develop, and test possible implicit bias mitigation strategies, we should focus on those that render salient important aspects of an individual's social identity, while limiting the effects of inaccurate stereotypes or pernicious associations.

Acknowledgements

I would like to thank Dan Kelly, Joshua Mugg, and an anonymous reviewer for comments on previous drafts of this paper. In addition, I'm grateful to the organizers (Jules Holdroyd, Alex Madva, Erin Beeghly) and participants of the *Bias in Context* workshop (2016) for their feedback and rich discussion.

[23] Although there are some scholars that think this is the ideal to which we ought aspire, I reject this view. For some empirical reasons for this rejection, see Kelly, Machery, and Mallon (2010).

References

Antony, L. 2016. Bias: Friend or foe? Reflections on saulish skepticism. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, 157-190. Oxford, UK: Oxford University Press.

Allport, Gordon W. 1954. *The nature of prejudice*. Cambridge, MA: Addison-Wesley Pub. Co.

Banaji, M.R. & Greenwald, A.G. 2013. *Blindspot: Hidden Biases of Good People*. New York, NY: Delacorte Press.

Brennan, S. 2016. The moral status of micro-inequities: In favor of institutional solutions. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*, 191-214. Oxford, UK: Oxford University Press.

Brownstein. Manuscript. *The implicit mind.*

Collins, P.H. 1990. *Black feminist thought*. Boston, MA: Unwin Hyman.

De Houwer, J. 2003. The extrinsic affective simon task. *Experimental Psychology*, 50(2), 77-85

Fricker, M. 2007. *Epistemic injustice.* Oxford, UK: Oxford University Press.

Fricker, M. 2015. Fault and no-fault responsibility for implicit prejudice—A space for epistemic 'agent-regret'. In Brady and Fricker (Eds.), *The Epistemic Life of Groups: Essays in the Epistemology of Collective*. Oxford, UK: Oxford University Press.

Froehlich, L., Martiny, S.E., Deaux, K., Goetz, T., & Mok, S.Y. 2016. Being smart or getting smarter: Implicit theory of intelligence moderates stereotype threat and stereotype lift effects. *British Journal of Social Psychology*, 55(3), 564-587.

Gendler, T.S. 2008. Alief and belief. *Journal of Philosophy*, October 2008, 634-663.

Gendler, T.S. 2011. On the epistemic costs of implicit bias. *Philosophic Studies: An International Journal for Philosophy in the Analytic Tradition*, 156(1), 33-63.

Glasgow, J. 2016. Alienation and responsibility. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*, 37-61. Oxford, UK: Oxford University Press.

Greenwald, A., D. McGhee, & J. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464-1480.

Greenwald, A., Poehlman, T., Uhlmann, E., & M. Banaji. 2009. Understanding and using the implicit association test: III meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17-41.

Hacking, I. 1995. The looping effects of human kinds. In D. Sperber, D. Premack, A.J. Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate*, 351-383.

Harding, S. 1991. *Whose science? Whose knowledge?* Ithaca, NY: Cornell University Press.

Holroyd, J. 2012. Responsibility for implicit bias. *Journal of Social Philosophy*, 43(3), 274–306.

Kelly, D., Machery, E., & Mallon, R. 2010. Race and racial cognition. In J. Doris and The Moral Psychology Research Group (Eds.), *The Moral Psychology Handbook*, 433-472. New York, NY: Oxford University Press.

Landemore, H. 2012. *Democratic reason: Politics, collective intelligence, and the rule of the many*. Princeton, NJ: Princeton University Press.

Levy, N. 2012. Consciousness, implicit attitudes, and moral responsibility. *Noûs*, 48, 21-40.

Machery, E. 2016. De-Freuding Implicit Attitudes. In M. Brownstein and J. Saul (Eds.), *Implicit Bias and Philosophy: Volume 1, Metaphysics and Epistemology*, 104-129. Oxford, UK: Oxford University Press.

Madva, A. 2016. Virtue, social knowledge, and implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, 191-215. Oxford, UK: Oxford University Press.

Mallon, R. 2016. *The construction of human kinds*. Oxford, UK: Oxford University Press.

Mandelbaum, E. 2016. Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, 50(3), 629-658.

Mills, C. 1998. Alternative epistemologies. In C. Mills, *Blackness Visible: Essays on Philosophy and Race*. Ithaca, NY: Cornell University Press.

Mills, C. 2000. But what are you *really*? The metaphysics of race. In A. Light, & N. Mechthild (Eds.), *Race, Class, and Community Identity: Radical Philosophy Today*, 23-51. Amherst, NY: Humanity Books.

Moss-Racusin, C., Dovidio, J., Brescoll, V., Graham, M., Hadnelsman, J. 2012. Science faculty's subtle gender biases favor moral students. *PNAS*, 109(41), 16395-16395.

Mugg, J. 2013. What *are* the costs of racism? A reply to Gendler. *Philosophic Studies*, 166, 217-229.

Mugg, J. Manuscript. How to deal with the tragic dilemma: An argument against the incommensurability thesis.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. 2002. Harvesting intergroup implicit attitudes and beliefs from a demonstration website. *Group Dynamics*, 6, 101–115.

Nosek, B. A., & Banaji, M. R. 2001. The go/no-go association task. *Social Cognition*, 19(6), 161-176.

Nosek, B., Greenwald, A., & M. Banaji. 2007. The implicit association test at age 7: A methodological and conceptual review. In J.A. Bargh (Ed.), *Automatic Processes in Social Thinking and Behavior*, 265-292. Philadelphia, PA: Psychology Press.

Payne, B.K., Cheng, C.M., Govorun, O., Stewart, B.D. 2005. An inkblot for attitudes: Affect misattribution as implicit measurement. Journal of Personality and Social Psychology, 89(3), 277–293

Rees, A. 2016. A virtue ethics response to implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*, 191-214. Oxford, UK: Oxford University Press.

Robertson, E. 2013. The epistemic value of diversity. *Journal of Philosophy of Education*, 47(2), 299-310.

Saul, J. 2012. Scepticism and implicit bias. *Disputatio*, 5(37), 243-263.

Shutts, K., Pemberton Roben, C.K, & Spelke, E.S. 2013. Children's use of social categories in thinking about people and social relationships. *Journal Of Cognition And Development*, 14(1), 35-62.

Smith, D. 1974. Women's perspective as a radical critique of sociology. *Sociological Inquiry*, 44, 7-13.

Stich, S. 2014. When democracy meets pluralism: Landemore's epistemic argument for democracy and the problem of value diversity. *Critical Review*, 26(1-2), 170-183.

Tetlock, P.F., Kristel, O., Elson, B., Green, M., Lerner, J. 2000. The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, 78(5), 853-870.

Washington, N. & Kelly, D. 2016. Who's responsible for this? Moral responsibility, externalism, and knowledge about implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*, 11-36. Oxford, UK: Oxford University Press.

Zheng, Robin. 2016. Attributability, accountability, and implicit bias. In M. Brownstein and J. Saul (Eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*, 62-89. New York, NY: Oxford University Press.