Article:

Holroyd, J. and Saul, J. (2018) Implicit bias and reform efforts in philosophy: a defence. Philosophical Topics, 46 (2). pp. 71-102. ISSN 0276-2080

PHILOSOPHICAL TOPICS
VOL. 46, NO. 2, FALL 2018

Implicit Bias and Reform Efforts in Philosophy: A Defence

Jules Holroyd
The University of Sheffield

Jennifer Saul
The University of Sheffield

ABSTRACT. This paper takes as its focus efforts to address particular aspects of sexist oppression and its intersections, in a particular field: it discusses reform efforts in philosophy. In recent years, there has been a growing international movement to change the way that our profession functions and is structured, in order to make it more welcoming for members of marginalized groups. One especially prominent and successful form of justification for these reform efforts has drawn on empirical data regarding implicit biases and their effects. Here, we address two concerns about these empirical data. First, critics have for some time argued that the studies drawn upon cannot give us an accurate picture of the workings of prejudice, because they ignore the intersectional nature of these phenomena. More recently, concerns have been raised about the empirical data supporting the nature and existence of implicit bias. Each of these concerns, but perhaps more commonly the latter, are thought by some to undermine reform efforts in philosophy. In this paper, we take a three-pronged approach to these claims. First, we show that the reforms can be motivated quite independently of the implicit bias data, and that many of these reforms are in fact very well suited to dealing with intersectional worries. Next, we show that in fact the empirical concerns about the implicit bias data are not nearly as problematic as some have thought. Finally, we argue that while the intersectional concerns are an immensely valuable criticism of early work on implicit bias, more recent work is starting to address these worries.

In recent years, there has been increasing attention in philosophy to issues of underrepresentation and of hostile and unwelcoming climates for members of underrepresented and marginalized groups. In addition to a very substantial and growing literature on this topic, there have been many initiatives designed to bring about change: the Collegium of Black Women Philosophers; Minorities and Philosophy; the blog What is it Like to be a Woman in Philosophy; the Gendered Conference Campaign; good practice guidance from the British Philosophical Association, the American Philosophical Association, and the Canadian Philosophical Association; the Site Visit Program of the American Philosophical Association; and so on. More and more departments and associations are setting up committees to reflect on and improve their practices, and more and more reforms designed to foster inclusion are being put in place.

One of the key justifications that has been used to argue for this attention, and for the reforms, has been empirical data on implicit bias.[1] Although this has been far from the only justification, it has received a large share of the attention—perhaps because this period has coincided with one in which implicit bias has been a burgeoning research topic in philosophy. However, the research from empirical psychology on implicit bias is far from uncontroversial. While indirect measures that are taken to access implicit bias have repeatedly shown patterns of bias in large samples, some concerns have been raised about whether this field of research yields adequate understandings to motivate reforms. First, worries have been raised about whether the research uses reliable tools to access mental constructs. Second, worries have been raised about whether whatever is measured by these empirical tools is a good predictor of behavior. Third, concerns have been raised about whether interventions to change implicit bias are the right place to focus our efforts. Moreover, concerns have been

[1] This has also been the case with respect to reforms outside of academia. To the extent that implicit bias has been relied upon to justify similar reforms in other contexts, the argument of our paper could be extended to those contexts also.

raised about the ability of research on implicit bias to address intersectional oppressions.

These concerns have been seized upon as evidence that reform efforts in philosophy are unmotivated or misguided, and have provided fuel for a backlash against efforts to diversify the profession (see e.g., Hermanson 2017, and countless online discussions).

In this paper, we first argue that reform efforts in philosophy are multiply motivated. Nearly all of these remain very well justified quite independently of results regarding implicit bias. The only ones that would be undermined are highly specific ones aimed exclusively at implicit bias, which have in general been some of the least widely adopted ones. So whatever is the case regarding implicit bias, philosophical reform efforts are far from misguided. Next, however, we demonstrate that there is no good reason to be so dismissive of justifications of reform efforts in terms of implicit bias—the data have not, after all, undermined these. They have shown that some particular claims made about implicit bias (including some of those occasionally made by one of us in the past) are in need of more careful phrasing. But this more careful phrasing does not undermine the justification for reforms. As we discuss reform efforts, we address whether they are able to take proper account of the intersectional nature of oppression; and we also show that empirical studies of implicit bias can, and increasingly do, take account of intersectionality.

1. UNDERREPRESENTATION/MARGINALIZATION IN PHILOSOPHY[2]

[2] A group needn't be underrepresented to be marginalized. Take, for example the case of LGBTQ+ philosophers, who may well—we don't know—not be underrepresented in philosophy, relative to the population. Even if this were so, they may still be marginalized, and so an appropriate focus of concern for philosophical reformers. (We are grateful to Kate Abramson for this point.)

In recent years, there has been an increasingly widespread recognition that philosophy is very white, male, middle class, and in countless other ways less diverse than it could be—and less diverse than many other fields, both in the Arts and Humanities and in the STEM subjects in which there has for some time been attention to issues of underrepresentation. According to our most recent data, women are 17 percent of full-time professional philosophers (where this does not include students) in the US (Norlock 2011), 24 percent in the UK (Beebee and Saul 2011), and 31 percent in Canada (Doucet and Beaulac 2013). But women are neither the only underrepresented group in philosophy nor the most underrepresented group. According to Botts et al. (2014), black PhD students and professional philosophers *combined* are just 1.32 percent of philosophers in the US.

The lack of diversity in philosophy varies across career stages and type of underrepresentation, and this variation in turn is affected by country. In the UK, women and men enter as undergraduate philosophy students in close to equal numbers (Beebee and Saul 2011), with a substantial drop-off at MA level, and further drops at subsequent stages. In the US, on the other hand, the largest drop-off point is at the point of deciding to major in philosophy (Paxton et al. 2012). A 2017 post by Eric Schwitzgebel based on data from the National Center for Education Statistics database found that Latino/Hispanic students and Asian undergraduate students were represented in philosophy at numbers comparable to their numbers in the population of undergraduates (as Schwitzgebel notes, Latinos/Hispanics are substantially underrepresented in the undergraduate population), but that blacks were substantially underrepresented relative to the student population (5 percent in philosophy versus 10 percent in the wider student population).[3] Botts et al.'s 2014 study of blacks in philosophy in the US suggests that this drops to less than 1 percent at PhD (2014, 237). Botts et al. found that there were 141 black people who were either employed by or PhD students in US philosophy departments; 35 percent of whom were women. Interestingly, Botts et al. also found that about half of black PhD students were women, a much higher percentage than among PhD students overall. A report by MAP[4] found that between 1995 and 2014, the percentage of undergraduate philosophy degrees awarded to members of traditionally underrepresented racial and ethnic groups rose from 9 percent to 17 percent; the percentage of master's degrees rose from 6.5 percent to 10 percent; and the percentage of PhD degrees from 2.7 percent to 8 percent. There have been particularly significant rises in the percentage of Hispanic students.[5] We don't have good data on the underrepresentation of disabled people, working-class people, or LGBTQ+ people.

We know, then, that philosophy is much less diverse than it would be if it were representative of the population (or even of the relevant university population, whether student or staff). We also know that there are many different and overlapping groups that are underrepresented, and that there are undoubtedly a variety of differing causes for the underrepresentation. Further, we know that—leaving aside issues of underrepresentation—members of marginalized groups report experiencing a hostile

[3] http://dailynous.com/2017/12/21/racial-diversity-philosophy-majors-guest-post-eric-schwitzgebel/?fbclid=IwAR1Xseo-NM_ng1J3VOvHYRZAP49Qhrv4KPHIMgVWDniMDGoVTdiPDvqpMqM.

[4] https://c.ymcdn.com/sites/apaonline.site-ym.com/resource/resmgr/data_on_profession/minorities_in_philosophy.pdf?fbc lid=IwAR3ApMc0Qn92-2N-yx4Fx5a0zD1sFfrXJ9YZRKbMeue6Zr8j54Md36UD_Bc.

[5] 'Hispanic' is the term used by the MAP report.

environment in philosophy.[6] Most of those who work on these issues are convinced that a wide range of causes are at work—some historical, some psychological, and some structural. A very plausible explanation is that something like Louise Antony's (2012) "perfect storm" model is broadly true, according to which a variety of different factors all affect philosophy, leading it to be a particularly severe site for underrepresentation. Implicit bias is just one of the factors that has been cited. However, it has been a rhetorically prominent factor. One reason for this is probably the excitement of philosophers about a deeply philosophically interesting notion that they have only relatively recently become aware of. A further reason has undoubtedly been the notable willingness of universities and departments to consider implicit bias as an explanation of the underrepresentation and marginalization issues that they are beginning to recognize. Implicit bias has, in short, probably played a large role in these discussions simply because both universities and philosophy departments have been especially receptive to this form of evidence (as opposed to, say, testimonial evidence from members of marginalized groups—which was available long before research on implicit bias came on the scene).[7]

2. IMPLICIT BIAS

What is implicit bias? While the concept has affinities with various ways of identifying subtle or unintended discrimination that predate the recent research programs of social

[6] See, for example, the descriptions from Anita Allen, in her interview with George Yancy, of the sorts of hostile environment she faces as a black woman in philosophy. https://www.nytimes.com/2018/06/18/opinion/black-women-in- philosophy.html. See also testimonies submitted to *What is it Like to be a Person of Color in Philosophy* blog (https://beingaphilosopherofcolor.wordpress.com) and some of those submitted to the *What is it Like to be a Woman in Philosophy* blog, such as this one: https://beingawomaninphilosophy.wordpress.com/2017/08/24/online-discussion-of-diversity-hires/.

[7] See Holroyd and Puddifoot, "Bias, Testimony, and Epistemic Injustice" (under review).

psychology, the term has been coined to capture a series of results drawn from studies involving indirect measures. Such measures do not rely on directly asking people to report on their attitudes, on the assumption that people will not report on some aspects of their cognition either because they are unaware of them, or because they are unwilling to reveal them. The most well-known and often used indirect measure is the Implicit Association Test (IAT), first developed in 1995 (Greenwald et al. 1998). This test measures the strength of associations between concepts or evaluations and social groups.[8] For example, a gender stereotype IAT can evaluate how strongly participants associate men and women with certain stereotypes to do with leadership or supporting roles, respectively (Dasgupta and Asgari 2004). A race evaluative IAT can evaluate how strongly white or black people are associated with positive or negative evaluations, respectively (Olson and Fazio 2006).[9] (IATs have been used to evaluate different stereotypical associations – different stereotypes associated with race and gender, as well as with other social categories, and likewise different sorts of evaluation. Other indirect measures include Affect Misattribution Procedure, Sequential Priming Measures, and Go/No-Go Attribution Tasks.[10]) Indirect measures have been used to identify different stereotypical and evaluative associations that participants hold relating to social groups including race, gender, age, disability, sexuality, size. The stereotypical or evaluative associations that are revealed on indirect measures such as the IAT are what are referred to as *implicit biases*. What is distinctive of implicit biases is that they

[8] The IAT has also been used to test associations between target concepts that are not to do with social groups, e.g., with brand associations.

[9] See Goff and Kahn (2013) and our discussion below for concerns that such studies are likely accessing gender stereotypes associated with *white women* and evaluations associated with *black men*, rather than with women and black people in general.

[10] For a useful summary of these different measures, see Michael Brownstein, “Implicit Bias,” *The Stanford Encyclopedia of Philosophy* (Spring 2017), Edward N. Zalta (ed.), especially section 1.2.

may not be available to individuals to report on; indeed, they may be in conflict with the agent's endorsed values and beliefs. Moreover, they are difficult to get rid of, and their influence on action can be difficult to control.

That our minds might encode such cognitions is troubling. More so the putative relationship with behavior. Studies have attempted to show that the presence of implicit biases (as measured by IATs or other indirect measures) correlates with behavioral tendencies. For example, a study examining implicit bias in medical practitioners found that to the extent that doctors had implicit biases against black people, the doctors tended not to prescribe treatment for black patients—patients whose described symptoms were the same as those of white patients for whom they did make prescription recommendations (Green et al. 2007). In a study on gender bias, Carlana (2018) found that the extent to which teachers had implicit gender biases—associating maths more strongly with men than women—correlated with classroom behaviors, such as expressions that affect girls' self-confidence, or fewer interactions with girl students—that led to differential outcomes for boys and girls in the classroom. To the extent that teachers had implicit gender biases about maths, girls' improvement was markedly less than that of boys in their class. Moreover, in a recent statistical analysis, Greenwald et al. suggest that implicit biases might have significant behavioral impact when their effects are considered across large groups (2015, 558). It is this putative relationship between implicit biases and behavior that has made compelling the suggestion that implicit biases might be part of the explanation for the persisting underrepresentation and marginalization of multiple groups in philosophy, and that measures, such as some of those described below, should be taken to guard against implicit biases affecting our interactions and evaluations.

3. PHILOSOPHICAL REFORM EFFORTS AND THEIR MANY JUSTIFICATIONS

Some efforts to reform the profession are very obviously not justified via implicit bias. The clearest case of this is reforms aimed at eliminating sexual harassment in philosophy. While it is entirely possible that implicit bias plays a role in perpetuating this problem (e.g., by leading people to find members of marginalized groups less credible than members of dominant groups), this is no part of the justification for doing something about sexual harassment. We will set aside cases like these and focus instead on reforms that seem more plausibly to depend for their justification on data about implicit bias.

3.1 ANONYMIZING: MARKING, HIRING, PUBLICATION

3.1.1 THE MANY REASONS FOR ANONYMIZING

Research on implicit bias is often used in arguing for anonymization. It is easy to see why: if we are likely to be influenced by implicit biases based on categories like gender and race—which can often be readily discerned from a name—it would be a good idea to keep us from seeing those names when we are making important decisions as we do in marking, hiring, and refereeing. Studies of CVs have shown a tendency to see the same CV as better with a typically white or male name (rather than a typically nonwhite or female name), and some of these have shown that this tendency correlates with implicit bias levels (see, e.g., Rooth 2010).[11][12]

[11] See also the study on CVs from Moss-Racusin et al. (2012)—but note the claim about gender bias there is based on the use of the Modern Sexism Scale, which is a self-report scale, albeit one that is more indirect than just asking people whether they hold sexist attitudes; however, it is obviously less indirect a

Anonymyzing (when it succeeds) would make it impossible for these biases to operate. Anonymous marking is simple enough to implement, and a near-universal practice in United Kingdom undergraduate marking (though obviously less successful in very small classes, or where markers also look at drafts).

Anonymous review at journals is fairly widely practiced, especially in philosophy. Fully anonymous appointment procedures for academic posts are harder, but some departments have anonymized at least one phase of the process—initial review of CVs, reading of writing samples, etc.[13] Anonymization works by concealing names, which may reveal not only gender but also sometimes nationality, racial group, or religion. This means that it is effective as a means of combatting (though obviously not eliminating) discrimination that might occur due to multiple dimensions of identity.

It might seem that the case for anonymization stands or falls with the literature on implicit biases. But this case in fact remains strong even if it is not based on implicit bias research.

measure than the IAT. As a result, Moss-Racusin et al. refer to subtle bias, rather than implicit bias—the extent to which such subtle bias is co-extensive with explicit or implicit bias is an open and interesting question.

[12] There have been criticisms of the idea that implicit gender bias works against women, in real world hiring situations. In particular Williams and Ceci (2015) argued that these preferences actually work in women's favor. However, their study focused exclusively on outstanding candidates. We have known for some time that biases are more likely to affect judgments in more borderline cases, and their work does nothing to undermine this. For further criticism of Williams and Ceci, see Brownstein (2015). Advancing a supplementary line of critique, Hermanson (2017) suggests that all cases of tenure-track hiring involve competition exclusively between outstanding candidates. But this is a puzzling suggestion. While it certainly is true that anyone who has acquired a philosophy PhD is outstandingly good at philosophy, it is still the case that (for example) the applicants with large numbers of publications in "top" journals are outstanding as compared to those with no or fewer publications. Insofar as hiring processes in philosophy include a range of candidates, it makes sense to suppose (contra Hermanson) that there will be categories of 'outstanding' and 'borderline' in any search. (We are actually skeptics about the idea of these top journals genuinely being better, but nonetheless we can recognize that CVs which contain publications in them are generally viewed as outstanding.)

[13] For example, Sheffield used anonymity to review CVs and cover letters in the longlisting process, then anonymously reviewed writing samples as part of shortlisting, as described here: https://feministphilosophers.wordpress.com/2014/07/11/on-conducting-a- job-search/.

Many studies that show differential evaluations of CVs (Steinpreis et al. 1999; Bertrand and Mullainathan 2004) are not studies of implicit bias.[14] They demonstrate that the same CV is judged differently depending on the name, but they do not make any effort to discern whether this is correlated with implicit bias. It is often assumed that the differential judgments observed in all of these studies are due to implicit biases. If, however, we came to doubt the influence of implicit bias on our judgments, we would still need to reckon with all of these studies. They would still show that our judgments are influenced by the perceived race and gender of the person we are judging. The most likely explanation of this, would be explicit bias. We would clearly also want to eliminate the influence of this, and anonymizing would still be a good technique to use.

Studies of bias and anonymization in marking, journal article review,[15] conference abstract submission[16] and grant review[17] are not entirely conclusive. This is an enormous and complex literature, so to get some sense of the complexity we will take a look at anonymization of marking. Bradley (1984, 1993) showed women at university to get higher marks under anonymous marking, which led to widespread assumptions that implicit marker bias was the cause of the initial lower marks. Hinton and Higson (2017) complicated this picture. Their large study of marking before and after the introduction of anonymous marking showed a small but significant narrowing of the achievement gaps

[14] This is not true of all of them, however. For example, Rooth (2010) found that callbacks of candidates with Arab-sounding names were negatively correlated with implicit stereotyping.

[15] Budden et al. (2008) appeared to show a substantial increase in women's publication rates after a journal instituted anonymous review. But this study has been criticized by Engqvist and Frommen (2008); Webb, O'Hara, and Freckleton (2008); and Whittaker (2008).

[16] Knobloch-Westerwick et al. (2013) asked subjects to review identical abstracts, accompanied by male and female names. They found a bias in favor of the "male" abstracts, particularly on stereotypically male topics.

[17] Wenneras and Wold (1997) found that female grant applicants needed to be more than twice as productive as male grant applicants to achieve success. However, subsequent studies have not shown this effect (see Lee et al. 2013 for a review of these). This may be due to measures put in place to improve gender equality after the original extremely well publicized results.

between groups after anonymity was introduced. However, this same narrowing was present for oral examinations, which were not anonymized. This casts doubt, at least in this case, on the thought that implicit bias in markers was responsible for the gap before the introduction of anonymity. The authors note that students' perception of greater procedural justice might instead be responsible for the change. In our view, this only enhances the case for anonymous marking: if there are marking biases to be blocked (implicit or explicit), anonymous marking is good. And even if there are not such biases, anonymous marking will reduce achievement gaps by creating a manifestly more procedurally just system.

Although much attention in this literature has focused on gender and race biases, there is very good evidence of other biases playing a significant role in academic hiring and review processes. In particular there is substantial evidence for nationality biases; and biases against non-native speakers of English.[18] There is also good evidence of affiliation bias, in which reviewers favor those with whom they have relationships.[19] To the extent that members of marginalized groups are less likely to be integrated into professional networks, this will work against such people. Finally prestige bias (favoring work that appears to have been written at a prestigious institution, people who work at such an institution, or people with other markers of prestige) is well established. Since members of marginalized groups may be less likely to have these markers of prestige, this will also disproportionately affect members of these groups.[20]

Perhaps the most important justification for anonymization, however, does not actually depend on any empirical results. Instead, it's the result of asking "*why not*

[18] For a good review of this evidence, see Lee et al. 2013.

[19] Again, Lee et al. 2013 provides an overview.

[20] Classic papers on this include Merton 1968 and Peters and Ceci 1982, but see Lee et al. 2013 for an overview and update.

anonymize?" If you think that judgments of quality of work or job candidate should be influenced just by the quality of that work or job candidate, why on earth would you want to have irrelevant information? We know enough by now, surely, to know that we are often influenced by things that we think shouldn't influence us. (And even if you think that you yourself are not influenced in this way, you undoubtedly know others who are.)[21] Removing the possibility of those influences, then, is surely a good thing. Unless one thinks that gender or race actually is relevant to these decisions, it is difficult to oppose anonymization.[22][23]

3.1.2 ANOYNMIZING AND INTERSECTIONALITY

While anonymization will deal (to some extent) with discrimination that may occur along multiple dimensions (e.g., race, gender, and class—insofar as these are identifiable on applications or publications), it is true that it is nonetheless limited in the ways captured by Crenshaw's 'basement' analogy. As she describes it, the basement contains people who are disadvantaged, sometimes along multiple dimensions. The ceiling is actually the floor of an upper room, in which privileged people reside. As Crenshaw writes: "in efforts to correct some aspects of domination, those above the ceiling admit from the basement those who can say that 'but for' the ceiling they too would be in the upper room" (1989, 151) . Access, however, is generally available only to those who, while

[21] Note also that evidence shows that confidence in one's own objectivity correlates with being more biased (Uhlmann and Cohen 2007).

[22] Of course, there may be some cases where gender or race is relevant to hiring decisions, such as in instances where one seeks to contextualize achievements or practice affirmative action. We discuss these cases below.

[23] There are, however, potential difficulties with anonymous marking when it comes to provision of certain sorts of feedback to students. Pre-submission comments on drafts, for example, compromise anonymity, as do other feedback methods that are intended to engage students in a dialogue with lecturers. Moreover, see Pitt and Winstone (2018) for the concern that anonymity might on some occasions affect student perceptions of the quality of relationship with their lecturers. It is possible that these concerns could give one reason not to anonymize under certain circumstances—perhaps with purely formative feedback—but difficult to see how this would justify more sweeping resistance to anonymity.

disadvantaged, are relatively privileged. Those who are multiply disadvantaged, experiencing oppression along many dimensions, are typically too far below to benefit from the access hatch to the upper room. Crenshaw's point is that to address some forms of intersectional oppression, more structural changes are needed than simply removing the discrimination that would not be experienced but for (typically) one aspect of identity (see also p. 144).[24] Anonymizing is the sort of reform that helps those who have, despite the many forces working against them, reached the same level of achievement as those from more privileged groups. But these will for the most part be those who are relatively privileged among marginalized group members. To fight intersectional oppression more fully requires structural changes that go well beyond anonymizing. Some of the reforms we outline below move toward more structural changes. Thus, we think it important to recognize that while anonymity is important, it is a limited strategy and other reforms are also needed to address exclusion.

3.2 OTHER HIRING REFORMS

There are a wide variety of reforms to hiring practices that have been suggested. We now turn to ones other than anonymity.

- Affirmative action: the most popularly suggested form of affirmative action is giving application materials from members of marginalized groups an additional look.[25] While concerns about implicit bias are one justification for this, there are many others. An obvious alternative justification would be explicit bias (though in that case, an additional look may be unlikely to remedy things). But there are also other reasons: Members of marginalized groups are likely to be less well mentored[26] so their materials may not be as polished as those of people who

[24] Note that Crenshaw's target is primarily the framework operative in antidiscrimination law, but she extends her critique to feminist thinkers more broadly also (1989, 150).
[25] Obviously, this is incompatible with a wholly anonymous procedure. But one could very well do this at the longlisting stage, and then ensure that writing samples are read anonymously.
[26] Moss-Racusin et al. (2012) found that participants reviewing the same CVs assigned to male or female names were inclined to offer more mentoring to male applicants.

have received better mentoring. They are likely to have encountered more obstacles, in the form of explicit bias and hostile environments, meaning the same level of achievement may have been more difficult for them than for members of overrepresented groups. Finally, there is some evidence that members of marginalized groups may tend to write on topics that are lower status.[27] Unless one thinks this status hierarchy is correct, it is a further reason to give these applications an additional look.

- Agreeing on hiring criteria in advance, considering specific issues separately, and taking time over decisions, while not putting excessive weight on any one part of the process: All of these are ways to move away from a quick gut reaction to a more considered view, arrived at in light of clear criteria. While these help with implicit bias, they are also helpful in making sure that decisions are carefully reasoned, and justifiable in light of agreed criteria. It is difficult to see how this could be resisted, even by an implicit bias skeptic.

- Reduce role of letters of reference. Research on letters of reference reveals some disturbing patterns, which have led some philosophical reformers to advocate reducing their role in hiring. However, this research is compatible with mechanisms other than implicit bias producing these patterns. Madera et al. (2009) showed that letters for women were more likely to use “communal” adjectives (e.g., ‘helpful’), while those for men were more likely to use “agentive” adjectives (e.g., ‘ambitious’), and that references with agentive adjectives were viewed more positively. Dutt et al. (2016) shows that women are only half as likely to receive excellent rather than good letters as measured by tone. They also found that—to the surprise of nobody who has ever been on a hiring committee—reference letters from those at institutions in the Americas are substantially longer than those from elsewhere.[28] The international variations go beyond this however. In some European countries, applicants write their own letters which are then merely signed by the letter ‘writer’, a feature of which readers outside these countries are generally unaware. And most British academics will readily accept that their national tendency to understatement makes their references less glowing than those from the US, though we have yet to see empirical data on this. Phenomena like these are not remedied by anonymization, which is why a case can be made for reducing or eliminating references’ role in hiring. Importantly, none of this case hinges on the presence or absence of implicit bias in letter writing. We might suppose that implicit bias contributes to the differential style of letter writing for men and women, but this is only one explanation. Indeed, none of the studies mentioned tests whether the biases and variations involved are implicit or explicit. But the evidence for these variations is substantial, and it counts strongly for reform.

[27] According to Botts et al. (2014), the top 5 areas for black philosophers are Africana, race, social and political, ethics, and continental. According to Haslanger (2009), the top 5 areas are feminist philosophy, applied ethics, normative ethics, social philosophy, and political philosophy.

[28] The Dutt study was international, so shows gendered patterns in a range of contexts. However, neither this study nor the Madera (2009) analyze how racial identity might interact with gender, so it is not clear how these patterns of differential reference writing interact.

- Snacks and breaks: decision making is better when people are well rested and well fed. This also reduces implicit bias, but surely better decisions and less cranky decision makers are justification enough.[29]

These strategies, like anonymity, will lead to better decision making, and thereby go some distance toward addressing intersectional concerns. But they are likewise somewhat limited in scope. It is important to note, however, that the first suggestion (about affirmative action) makes room for the thought that members of marginalized groups may appear to be less well qualified if we simply apply the criteria we have already applied. It recognizes that if people have been encouraged less, or mentored less, or encountered more obstacles, they will appear to be less promising than they actually are. This suggestion, then, is not simply one of leveling the playing field. It is also a call to acknowledge the many ways in which the playing field is *not* level—and this has the potential to better address some of these concerns.

3.3 RUNNING DISCUSSIONS DIFFERENTLY

In many places, philosophical discussions (both in classrooms and at seminars and conferences) are understood as battles of sorts. Very commonly, the goal at departmental seminars has been (sometimes implicitly, sometimes explicitly) to destroy the speaker. Also commonly, a few high-status voices tend to dominate, with multipart questions and endless follow-ups. These sorts of discussion norms have been the subject of critiques for some time (Moulton 1983; Rooney 2009; Beebee 2013; Saul 2013). Reforms commonly suggested include (a) encouraging the idea that the goal

[29] One of the most compelling pieces of evidence in favor of snack breaks is provided by S. Danziger et al. (2011), suggesting that judges make harsher decisions immediately before lunch and snack breaks. This study has been criticized: Glöckner (2016) has suggested that part of this variation is susceptible to an alternative explanation, but only part of it.

should be philosophical progress rather than victory; (b) instituting a one-question per question rule; (c) giving precedence to those who have not yet spoken. In teaching, a common suggestion is to have explicit discussions with students, in which they propose and discuss rules for productive discussion.

Implicit bias sometimes plays a role in justifying these reforms—after all, it may contribute to overlooking voices from marginalized groups or to favoring the high status. It can also lead to some men's verbal aggression being viewed favorably and women's unfavorably. (See, for example, Antony [2012] on the 'double bind' women face because arguing like a "good philosopher" clashes with norms for being a good woman.) Moreover, note that certain behaviors from white men may be tolerated but may not be from black men, whose behavior is instead regarded as intolerably aggressive (Duncan 1976). Evidence from various workplaces (including universities) suggests black women's assertive behavior may also be penalized in distinctive ways, with the 'angry black woman' stereotype always looming (Williams 2014, 201–2). So not only are the 'battle' norms not conducive to discussion; they also may be applied with double standards to perpetuate further exclusions. Evidence of implicit bias, then, is not needed to justify reforms: if one's goal is to do good philosophy, destroying every speaker is a perverse way of going about it. Encouraging, and making it comfortable, for more people to speak makes discussions more interesting and productive. And students benefit greatly from reflecting on how to have useful discussions.

Working together toward norms that are maximally inclusive—especially since discussion is such an integral part of our discipline—moves toward the sort of structural change we emphasized as important above. (However, it is worth noting that this change probably only affects those who are already 'in' the discipline—exposed to or at some level engaged in its discussions. Where there is already underrepresentation at the early stages of disciplinary participation, it is not clear that this strategy will remedy that.)

3.4 WORK HARDER TO NOTICE CONTRIBUTIONS

There are lots of data, both empirical and anecdotal, showing that contributions from members of marginalized groups often go unnoticed. Famously, women in meetings have their points ignored until they are repeated by (and often attributed to) a man (the phenomenon of ‘hepeating’ as it has become known. See also McClean et al. 2018 for research showing the different reception that men and women’s contributions receive). Excellent work by members of marginalized groups rarely makes it into introductory anthologies or onto syllabi (Thompson et al. 2016). Women are cited less than men (Healey 2015; Maliniak et al. 2013). And, of course, there are still a lot of conferences with all-male lineups of invited speakers[30]; more yet with all white lineups.

It’s commonly thought that implicit bias is an important cause of these phenomena. But we don’t really know this—as far as we know, there haven’t been any studies showing (for example) a correlation between the phenomena just mentioned and IAT scores. All we know is that these forms of overlooking exist. If it were to turn out that implicit bias isn’t a causal factor in these phenomena, it would still be the case that these forms of overlooking exist—but we would need other explanations for them. These are readily available in the form of explicit bias, both against social groups and against particular topics or approaches. But there are also historical explanations, for

[30] For a few of these, and discussion of these issues, see the Gendered Conference Campaign: https://feministphilosophers.wordpress.com/gendered-conference-campaign/.

example, for not putting women in introductory anthologies—women have not traditionally been a part of the canon,[32] and introductory anthologies often simply draw on the inherited canon.

Even if we became convinced, then, that implicit bias did not cause these phenomena, we would still believe that the phenonema exist. And we would still want to do something about it—it is better not to overlook good work, and where there's a systematic tendency to overlook good work from members of particular groups that's unjust. People deserve credit for their ideas, so trying hard to notice contributions in meetings and to cite relevant work should be uncontroversial proposals. Moreover, there is reason to believe that seeing members of marginalized groups taken seriously as experts helps members of those groups to succeed, and this research in no way depends on results regarding implicit bias.[33] The solution suggested—try harder to notice contributions—will be equally effective even if the not-noticing stems from a cause other than implicit bias.

Again, these kinds of changes—to what and who is included in debates and on curricula—moves toward the sorts of structural changes that are needed to be more thoroughly inclusive; though in order to do this, the inclusion needs to be more than merely 'tokenistic'.

[32] Though see O'Neill (1997) on the extent to which early modern women were a part of the mainstream in their own time, only to be dropped from the canon later. This may well have been true for other groups and times.

[33] Dasgupta (2011) argues that members of underrepresented groups who are experts may serve as a kind of "stereotype inoculation" that helps members of those groups to succeed. Her focus is on *contact* with those experts (as lecturers/professors), so it does not provide conclusive support for the idea that encountering the work of these experts will have the same effect. However, it makes this hypothesis a plausible one.

3.5 DIVERSIFYING READING LISTS, CONFERENCE LINEUPS, ANTHOLOGIES, VISITING SPEAKER SERIES

This point has already been touched on above, since leaving people out of reading lists and conference lineups is a way of overlooking their contributions. But the solution of actively working to diversify reading lists, conference lineups, anthologies, and speaker series is controversial enough to merit separate discussion. Moreover, there are multiple kinds of diversity that may be at stake. Here we touch on just two.

3.5.1 DEMOGRAPHIC DIVERSITY

The push for demographic diversity is for more members of marginalized groups. One of the efforts in this domain is the Gendered Conference Campaign, which calls attention to conferences with all-male invited speakers in the hope of encouraging people to avoid this. But there are also plenty of other efforts along these lines. The BPA/SWIP guidelines, for example, call for diversifying reading lists (their focus is specifically on gender, but some departments have adopted a broader version), and for trying to avoid all-male lineups of speakers. There are now some wonderful online resources to promote the diversification of invited speakers and reading lists, such as the Diversity Reading List (https://diversityreadinglist.org) and the UP Directory (https://updirectory.apaonline.org). Although implicit bias is often cited as an explanation for the current lack of diversity, in reality the problem is the lack of diversity—and the problem remains whether it is caused by implicit bias, explicit bias, structural factors, or all of these things. Implicit bias comes into justifications that focus on the potential effects of diversifying: the presence of counterstereotypical exemplars (members of the stereotyped group who don’t fit the stereotype) can help to reduce implicit bias. But there is also ample evidence, not dependent on implicit bias research, that diversifying demographically helps to increase the sense of belonging of people from marginalized groups (Dasgupta 2011); and that reaching a critical mass of people from underrepresented groups can improve the accuracy of judgments (Stewart and Valian 2018, ch. 5). Moreover, once we acknowledge the intersectional nature of oppression, it is clear that just adding *women* to reading lists is insufficient. Instead, it becomes vital to consider all the dimensions of demographic diversity, and to work to improve on as many as are reasonably possible.

3.5.2 DIVERSITY OF SUBJECT MATTER

Another important reform is to diversify in terms of subject matter: instead of simply teaching canonical issues, this line of argument goes, it is important to teach issues such as gender, race, class, and disability: ones that have traditionally been neglected in philosophy. This reform is generally *not* justified in terms of implicit bias, though a case could be made that implicit bias has played a role in the neglect of these issues—for example, by causing members of the profession to take those who raise them (generally members of marginalized groups) less seriously. Generally, this reform is justified on straightforwardly philosophical grounds.

Take, for example, the remarkable absence of discussions of racial justice in the mainstream philosophical literature on justice, until very recently. Proponents of nonideal theory like Charles Mills (2005) argue that attempting to theorize justice without attention to real-world injustice leads to serious errors and omissions. And these go unremarked and unnoticed, because of lack of attention to these issues. He notes that not only does Rawls fail to discuss racial injustice except in passing, but that until very recently the vast literature on Rawls also ignored this. Moreover, he notes, Nozick's immensely influential libertarian theory has startling consequences of modern-day property rights if one considers the immense historical violations of Native American property rights—but this too went largely unremarked for some time. In some instances, *what* has been left out is not unrelated to *who* has been left out, as Charles Mills suggests; only from a position of white ignorance could ignoring racial injustice be possible. Once this is acknowledged, it is clear that ignoring issues of racial injustice impoverishes philosophical discussion of central topics like justice. Once more, this argument has nothing to do with implicit bias.[34]

3.6 REJECT "GENIUS" IDEAL

A recently popular suggestion in philosophy has been to work to rid ourselves of the "genius" ideal, according to which philosophical insight and success comes from a fixed innate ability. Some (such as Schwitzgebel 2010 and Saul 2013) have suggested that philosophers' views about who "seems smart" are influenced by implicit biases. But the main empirical evidence for rejecting the genius ideal is the work of Leslie et al. (2015), which examines the extent to which members of a field take success to be due to an innate ability, and the extent to which women and black people are underrepresented. They found a significant correlation between these: the fields in which success was most thought to be due to innate ability were also the fields in which women and black people were most underrepresented.[35]

[34] Note that the call to include diverse subject matter is sometimes presented as a matter of 'intellectual diversity'. For concerns with this framing, see Dotson 2018.

[35] It isn't clear, from their analyses and report, how the intersection of gender and race applies here, that is, whether beliefs about the causes of success interact with the representation of white women or black women, or Asian American women—the other racial group in the analysis—to the same degree or differentially.

3.7 MENTORING SCHEMES, SUMMER SCHOOLS

As noted a few times above, members of marginalized groups will often have received less mentoring and support than others. Several important sorts of programs have come into existence to attempt to address this.

- Programs like PIKSI and Rutgers Summer School for Diversity in Philosophy offer summer schools for promising undergraduate members of marginalized groups who are considering postgraduate work in philosophy. They get mentoring from both senior members of the profession and current PhD students, and meet other marginalized students considering philosophy.

- Mentoring workshops like the one-off SWIP UK mentoring workshop, or the regular Mentoring workshops run by the Mentoring Project in the US, offer early career women the mentoring support of senior women, along with the chance to spend time with other women of their own cohort.

- SWIP UK's mentoring program works differently, pairing junior women (from postgraduates on) with senior members of the profession of any gender for mentoring support, tailored to the particular requests from the junior women.

These programs are not justified by invoking implicit bias, even though implicit bias may of course be a factor in the lack of mentoring that members of marginalized groups receive. Rather, they are initially justified by noting that members of these groups receive less mentoring and want more mentoring; and their continuation is justified by the testimonies of those who have participated. These programs—especially ones that focus on multiple intersectional categories like PIKSI and Rutgers—are especially well suited to addressing the problems pointed to by Crenshaw's basement analogy.

3.8 INDIVIDUAL BIAS-REDUCTION TECHNIQUES

Some reform proposals have specifically focused on reducing implicit biases. These draw on research from social psychology, where interventions to reduce implicit bias have been studied. As such, these strategies all take the form of 'cognitive fixes': things that individuals might do in order to change or reduce the extent to which they harbor implicit biases (with the hope that reducing bias is then reflected in behavior—an issue we take up in section 4.3 below). One such intervention focuses on 'implementation intentions': intentions to think counterstereotypical thoughts cued to particular contexts (such as 'if I am in a talk by a woman, I will think 'competent').[36] Implicit bias training developed by Devine and colleagues (reported on in their 2012) asks participants to adopt a suite of strategies to try to reduce their own biases: stereotype replacement; counterstereotypic imagining; individuating (focusing on specific features, rather than group membership); perspective-taking (imagining what it is like for others); and increasing opportunities for contact with stigmatized group members.

Adopting such strategies seemed somewhat successful: participants who did so experienced a reduction in implicit bias on later IATs. They also showed greater concern about discrimination, and increased awareness of their own propensity for bias, at two months' follow-up. This sounds like a good outcome—though note that a 2017 replication of the study found similarly *except*, crucially, that there was little reduction in implicit bias (Forscher et al. 2017a). Still, raising awareness of propensity for bias and increasing concern for discrimination is not insignificant—after all, the greater an individual's concern, the more likely she will adopt some of the reform strategies outlined above, one might think. However, it is not obvious that there is robust empirical

[36] This strategy draws on research from Stewart and Payne (2008), who looked at implementation intentions concerned with stereotyping of black people (as it turns out, in this context the stereotypes target black men). The implementation intentions they focused on addressed 'danger' stereotypes: "Whenever I see a black face on the screen, I will think the word 'safe'" (1336).

support for adopting individual bias reduction strategies. Nor is it obvious that these are where energies are best placed—as we shall see in the following sections.[37]

3.9 IMPLICIT BIAS TRAINING

A very widely adopted reform effort—not just within philosophy—is to institute implicit bias training. Such training typically constitutes a workshop aimed at raising awareness about implicit bias, and—to greater or lesser degrees—attention to what might be done to combat biases. It might seem obvious that the usefulness of this intervention depends on what the research about implicit bias shows, and it is certainly true that there is some relationship here: if we became convinced that implicit bias did not play any important role in marginalization and underrepresentation, or—at a further extreme—that it did not exist, it would be bizarre indeed to endorse something called 'implicit bias training'. However, it is worth noting that there are very many different kinds of things that go under the heading of 'implicit bias training'. Some offer no suggestions at all for how to reduce the manifestation of implicit bias. As we'll see in a later section, these trainings are by far the most problematic, as there is good empirical evidence that this kind of training should not be done. Others focus on individual bias reduction strategies, as discussed in section 3.8 (see, e.g., Devine et al. 2012). But the sort of training we both favor—and offer when we run training sessions—is very different from these. We favor implicit bias training on which implicit bias offers a useful way in to discussing the structural, institutional, and cultural factors that give rise to and perpetuate these biases. And although we do include some discussion of individual-

[37] Though see Madva 2017 for a robust defense of the value of institutionally sponsored programs of individual bias-reduction training.

level remedies, we emphasize the importance of the sorts of much broader sorts of reforms we have been discussing in this portion of the paper.[38] These reforms are not justified only in terms of implicit bias, so training that presents and argues for these reforms is multiply justified. Empirical concerns about implicit bias would certainly provide good reason for altering these training sessions (especially with respect to their names!), but much of the training would remain very well justified.

4. THE SOURCES OF SKEPTICISM

So far, we have argued that there remain good reasons to undertake various reforms irrespective of whether one relies on research from empirical psychology about implicit bias. One reason for which some may be reluctant to draw on such research is that there have recently been claims advanced that challenge the robustness of this research. Some of the concerns with research on implicit bias in fact targets just one of the methods that has been used to access implicit attitudes: the Implicit Association Test (IAT).[39] This test measures how quickly people are able to make stereotype congruent or incongruent categorizations. (For example, categorizing *men* and *leadership* would be stereotype congruent; *women* and *leadership*, or *men* and *supporter* would be stereotype incongruent categorizations.) If, as is frequently found, people are able to respond more quickly in making stereotype congruent categorizations, it is inferred that those notions are more accessible, and hence more likely to be activated. When activated, biases may have an influencing role on judgment or behavior. Indeed, some studies have tried to show that to the extent that one has

[38] For a fuller discussion of the ways that implicit bias training can be used to motivate structural and cultural reforms, see Saul (forthcoming).

[39] This means that other measures used to access implicit biases remain unchallenged by these claims.

implicit biases of some kind, one is more likely to behave in discriminatory ways (Green et al. 2007). However, various critiques have been raised in relation to the research program using indirect measures such as the IAT. We set out, and defend against, these critiques below.[40] So, we argue below, one should feel free to justify reforms by appealing to implicit bias; moreover, there can be some important gains from doing so. But since the reforms are not solely dependent on this research, this is not the only source of support for them.

4.1 TEST-RETEST RELIABILITY

Doubts have been raised about the reliability of the IAT. In short the worry is this: if the test is accessing some stable mental construct, then we would expect the test to yield the same results when people take it over time (unless there is reason to suppose some intervention has altered that stable attitude). But, people's IAT 'scores' can vary from occasion to occasion on which they take it—there is low test-retest reliability. This suggests, it has been argued, either that the IAT does not measure some stable mental construct, or that, to the extent that there is such a construct, this tool is not one that reliably accesses it.

How strong a challenge is this? First, to the extent that there are other indirect measures that access implicit attitudes,[41] it appears that the best interpretation of the challenge is to suppose that it targets the reliability of the tool (rather than the existence of any stable mental construct). So, is the IAT an unreliable tool? It is true that if you

[40] For other responses to critiques of the IAT, see Jost (2018), Gawronski (2018), Brownstein, M., A. Madva, and B. Gawronski (ms), and Brownstein (ms).

[41] For example, consider an IAT that accesses associations between black and weapons. Other indirect measures have revealed these associations also, in particular sequential priming tasks (priming with a black rather than white face increases the frequency at which ambiguous objects are identified as guns) (Eberhardt et al. 2004; Payne 2001).

take a test one day, then take it again a week later, you might not get the same score. However, this is completely unsurprising, given that what is measured are fine-grained response times, and we know that these are influenced by lots of situational factors—how tired you are, recent interactions, other beliefs and goals activated at the time you take the test. It is rare that all of these are held fixed across different occasions on which individuals participate in these measures (especially where participation is online, and the experimenters have no control over background conditions in which individuals participate). This fact has been known for some time, and it is a key part of the basis for bias-reduction techniques like those discussed in 3.6.

What conclusions, then, can be drawn from this? First, note that the performance of the IAT is not anomalous here. As Jost notes, "the IAT shows higher (within persons) test-retest reliability than other measures commonly used in psychological research, including Stroop and priming tasks" (2018, 2). Second, researchers who use the IAT are generally at pains to emphasize that the IAT is not diagnostic—it won't tell us what an individual's *real level of bias* is: indeed, the strength of association can change.[42] However, consider that results on large projects, such as those of Project Implicit, in which hundreds of thousands of people have participated, show certain patterns emerging notwithstanding variation in individual scores. People tend to have stronger negative associations with black rather than white people; people tend to associate men more strongly with science or leadership than they do women.[43] These broad patterns of results *include* those from individuals whose scores have changed from time to time

[42] It should be admitted, however, that those discussing the IAT sometimes write as though *individuals* have particular levels of implicit bias, rather than *individuals at times or individuals in particular circumstances*. Jennifer Saul is among those who have made this error, which she now regrets.

[43] Again, we note this with the caveat about the gendered assumptions about race, and racialized assumptions about gender, likely to be operative in these studies (see note 9 above and discussion below).

that the IATs have been taken. Notwithstanding this individual variation, the pattern of bias emerges, and this is where the real import of the research finding is. If large numbers of people all have even slight biases that tend in one direction, this could have a significant impact (see Greenwald et al. 2015 for a simulation to this effect). It has led Payne et al. (2017) to develop the Bias of Crowds model. They note that while implicit biases vary across time for individuals, they tend to stay stable across locations such as countries or cities, and to be correlated with levels of discrimination. According to the Bias of Crowds model, we should view implicit bias measurements more as measurements of situations than of individuals. Others propose a less radical model which nonetheless accounts for the 'person in the situation', or emphasizes the significance of 'contextual factors' (see Jost 2018 and Gawronski 2018, respectively) in the measurement of implicit bias.[44] These models, which emphasize the importance of situational context, serve very well to justify many of the reforms discussed in section 3. In particular, they motivate the kinds of reforms to institutional procedures that we identified—more inclusive curricula, events, discussions, and departments, affirmative action in hiring, say—rather than any particular individual bias reduction strategy in the form of a 'cognitive fix'. The focus on creating better environments is a way of taking seriously the ways in which the expression of biases seem to be heavily affected by context. These more structural changes acknowledge widespread nature of the problem and its environmental causes, and do not suppose that in each instance in which we might find bias there will also be an effective 'cognitive fix'.[45]

[44] Gawronski: "the available evidence suggests that contextual factors determine virtually every finding with implicit measures, including 1) their overall scores, 2) their temporal stability, 3) their prediction of future behavior, and 4) the effectiveness of interventions" (2018, 11).

[45] It should be also noted that the standards for test-retest reliability are not uncontroversial, and critics often invoke a particularly high, and we think implausible, standard.

4.2 LACK OF PREDICTIVE VALIDITY

A stronger source of concern focuses on the predictive validity of the IAT and other indirect measures. Predictive validity concerns the extent to which how one performs on a measure (such as the IAT) predicts how one will perform on other measures—of particular relevance here are behavioral measures such as interactions, judgments, or other tasks that manifest attitudes. Recent meta- analyses have suggested that even if the IAT (or other indirect measures) does reliably reveal some stable mental construct, that mental construct does not itself correlate with any particular behaviors that we might care about, such as discriminatory behaviors (Oswald et al. 2013). We think, however, that these concerns are misguided. One reason for this is that there are serious problems with the methodology of the meta-analysis; the other is that low predictive ability is not actually as worrying as others have made it out to be.

The methodological criticisms are those made by Greenwald et al. in their response to the Oswald meta-analysis. Greenwald et al. point out that the inclusion criteria used by Oswald et al.—namely, the grounds for including a study in the meta-analysis—are seriously problematic. Oswald et al. looked at the relationship between bias and behavior on 46 studies concerning racial and ethnic biases. However, while some of these studies predicted behaviors, others of them were testing specifically to show that the results of the indirect measure *did not* predict a certain kind of behavior. For example, Oswald et al. include studies from Amodio and Devine (2006), who predict that an IAT that measures associations between race and stereotypes of intellectual abilities will predict judgments of competence, but that an evaluative IAT would not. Evaluative IATs, which look at positive/negative attitudes toward black and white individuals more generally, are instead predicted to correlate with warmth behaviors, rather than judgments of competence. It is a mistake to include the relationship between evaluative IATs and judgments of competence (or stereotype IATs and warmth behaviors) in a meta-analysis, since these studies were designed precisely to show that the measure *does not* correlate with some behaviors (while it does correlate with others). Yet Oswald et al. included such studies in their meta-analysis. When Greenwald et al. included in the analysis only studies where a relationship with behavior was expected, they found that predictive validity was considerably higher (.216, rather than .15). In other words: the predictive validity is much higher when the meta-analysis is confined to *what was actually predicted*.

Still, while higher than the original meta-analysis, psychologists would describe the predictive validity as 'low'. Is this a worry? Greenwald et al. argue not—that small effect sizes can still be consequential, especially if the effects are found in large numbers of people. They use statistical modeling to demonstrate this. This is a very important point. However, we want to advance a different strategy to emphasize the importance of even low predictive validity.[46]

Our response starts by noting that few among us would think that the prevalence of explicitly sexist or racist beliefs are unproblematic, even if we learned that in particular cases they often don't affect behavior. Instead, we'd quickly acknowledge that of course many factors have an effect on what behaviors people will manifest, and that these may mean that explicit beliefs don't cause the behaviors we'd expect. And this is in fact just what we find. Across these same studies the predictive validity of *explicit* beliefs is also

[46] See also Gawronski's response to concerns about predictive validity: he suggests that there is no basis for expecting a strong relationship between implicit bias and behavior (2018, 9).

very low (.12). And this is not different from the general pattern for explicit beliefs—models for predicting behavior have typically moved beyond appeal just to beliefs, since beliefs are seen to be poor predictors of behavior in a range of domains. For example, in the health sciences, where predicting healthy or risky behavior is important, various studies show the poverty of beliefs as predictors of behavior (see Abraham and Sheeran 2005 for an overview).[47]

This casts a different light on the low predictive validity of implicit biases—it is low, but so is the predictive validity of *explicit beliefs*, whose importance few would (or should) doubt. Even if beliefs are not (in some domains) good predictors of behavior, we still have reason to care about them. First, we might simply maintain that it is wrong or bad to have problematic (false, misleading, inaccurate, stereotyping …) cognitions. We might think this about explicit beliefs and implicit cognitions. Second, we might think that a further thing that is problematic about such cognitions is that they present the possibility or risk of being expressed or acted on. We might want to remove or reduce that risk, no matter how small. Moreover, the level of risk is going to be strongly influenced by contextual factors such as the opportunity to act on such cognitions, the encouragement to act on such cognitions, and the social and practical consequences of doing so (see, e.g., Duguid and Thomas-Hunt [2015] for evidence that people may be more likely to express bias when they believe it is normal to do so). This gives us an excellent reason to care about creating environments that will reduce the risk of these cognitions being expressed or acted on—which is one of the goals of the professional

[47] In fact, some beliefs at least intuitively are better predictors of behavior. For example, my belief that the shop has closed will predict that I won't take a detour to the shop on the way home; my belief that I will meet Alix at 10 a.m. is a good predictor of my turning up at the agreed location to meet her. Some beliefs appear to be good predictors of behavior, others bad; it is an interesting question, but one beyond the scope of this paper, what characterizes beliefs of each sort, and explains their different relationship to behavior.

reforms discussed above. So these two reasons mean that we should care about implicit cognitions, even if the extent to which they predict behaviors is low. A third reason we might appeal to for caring about explicit beliefs is that we might assume that changes in beliefs will be accompanied by changes in behavior. However, this consideration takes us directly to the third concern to be addressed in the following section.

Finally, notice that the concern about predictive validity is focused on the extent to which the measures are predictive of individual behavior. But, as we have noted above, the best recent models of implicit bias focus on its effects as a group-level phenomenon (e.g. Payne et al.'s [2017] 'Bias of Crowds' model). As they point out, the relationship between patterns of implicit bias at group level, and the discriminatory behavior of that population, is more robust (7–8, 19).

To summarize the case so far: the predictive validity of implicit biases appears not to be as low as Oswald et al. initially claimed. But in any case, we have reason to care about implicit biases even if they are poor predictors of behavior, just as we may still care about explicit beliefs that are similarly poor predictors of behavior. Moreover, understood as a group-level phenomenon, the relationship with behavior is more robust.

4.3 INEFFICACY OF BIAS CHANGE

A further worry is that focusing attention on implicit biases is simply the wrong place to direct our efforts. In particular, this concern has force when one considers the results of a meta-analysis showing that changes in behavior did not appear to be mediated by changes in bias (Forscher et al. 2017b).[48] In that case, why bother trying to change biases at all?

We think that it is right to embrace this concern about efforts to simply *change biases*. On the one hand, it is completely unsurprising. On the other, explicitly recognizing and acknowledging it helps us see where we might best place our efforts. In order to understand the force of this concern and its implications, it is again worth thinking about the role and relevance of changing cognitions in other contexts. Consider the health sciences, where much research focuses (inter alia) on how to get people to stick to their exercise regimes, take their medicines, quit smoking, use condoms, etc. It is well acknowledged that changing isolated cognitions alone is insufficient for robust behavioral change. Simply getting individuals to believe in, say, the importance of adhering to an exercise regime is notoriously inadequate for getting people to stick to it (an insight that no doubt resonates with much personal experience!) (see, e.g., Abraham and Sheeran 2005). Instead, strategies that involve putting in place environmental cues or props to help one stick to the exercise plan have better success. For example, people who signed up for text message reminders (‘time to exercise!’) did far better at sticking by their exercise regimes than those who simply thought it would be good to exercise (Prestwich et al. 2009). The upshot is: it is well known that changing behavior does not depend solely on changing cognitions, but on more creative strategies that get us to do what we otherwise can’t quite manage. Environmental and institutional interventions seem to be the most effective, as they help us to create

[48] Though note that, insofar as the changes involve a reduction in, rather than elimination of, implicit bias, it is not surprising that significant behavioral changes are not also seen.

situations conducive to getting us to act in accordance with our all-things-considered values.

All this is to say: if studies show that changing behavior is not brought about just by changing biases, this just gives us reason to embrace the conclusion that our attention should not focus primarily or solely on *changing or reducing* implicit biases. This does not mean that implicit biases are not an important problem to address, but rather that addressing them shouldn't (alone, or even primarily) take the form of reducing implicit biases.[49] This is a welcome conclusion, since—as noted in section 4.3 above—there is little robust evidence on quite how to reduce biases in the mid- to long term (Forscher et al. 2017a, 2017b). And, indeed, with the exception of the 'cognitive fixes' discussed in 3.8, the professional reforms suggested do not focus on changing or reducing biases. Instead, in the cases where they are related at all to implicit biases, they focus on changing our environments.

As it turns out, then, reform strategies such as those considered in sections 3.1–3.7 are not only multiply justified—good practice independently of concerns about implicit bias—they also look like the best strategies for responding to concerns about implicit bias.

4.4 CONCERNS SPECIFICALLY ABOUT IMPLICIT BIAS TRAINING

There is research that has been taken to cast doubt on the efficacy of implicit bias training; indeed some of this research has been taken to show that implicit bias training is counterproductive. We think this is absolutely right for *some* forms of implicit bias

[49] For example, changing biases might be a necessary but insufficient part of changing behavior. This is consistent with the claim that we should not focus just on changing biases, but on changing environmental contexts—one of the downstream effects of the latter could be (more effective) bias change.

training. But we think there is no evidence against other forms of implicit bias training, and also that there is good reason to think they could be helpful. Let's begin by looking at a bit of this research.

Critics of implicit bias training often cite Duguid and Thomas-Hunt (2015). The key finding that is taken to undermine implicit bias training is presence of a moral licensing effect from learning that implicit bias is widespread: those who are told just this become more likely to make stereotypical judgments.[50] If this is right—and there's no reason to think that it isn't—then any implicit bias training that simply tells people that implicit bias training is widespread is counterproductive. We agree wholeheartedly with this, and we consider it an enormous mistake to offer implicit bias training of this kind.[51] Anecdotal reports suggest that this sort of training is quite widespread, and if that's right then it is something that really needs to change.[52]

Importantly, however, not all training is like this. Some training specifically focuses on the importance of combatting implicit biases. For example, Devine et al. (2012) tested training that offered participants a suite of individual techniques for reducing implicit bias and found this to be effective in reducing bias, and increasing reported levels of concern about implicit bias. [53] However, a replication of this study with a much larger sample failed to show the training to have an impact on implicit biases (Forscher et al. 2017a). A more recent study (Carnes et al. 2015) has also raised some questions

[50] Interestingly, the same study shows that if people are instead told that most people attempt to overcome their implicit biases, they are less likely to make stereotypical judgments.
[51] Saul (forthcoming) argues that this sort of training is analogous to a driver education course that merely teaches about the prevalence of accidents, without offering any methods for avoiding them.
[52] Anecdotal reports also suggest the prevalence of training in which people are told that they can overcome implicit bias by being objective. Uhlman and Cohen (2007) show that this is also likely to increase manifestation of implicit bias. There are also many reports of implicit bias training in which people are given guidance on how to overcome explicit bias, but no guidance on how to overcome implicit bias—but told that they have been trained in overcoming implicit bias. We would expect this to be counterproductive as well.
[53] Also see Madva (2017) for an argument in favor of teaching individual bias-reduction techniques.

about the efficacy of these individual bias-reduction techniques. At this point, there is conflicting evidence about the efficacy of well-run training that focuses on individual bias-reduction techniques. However, this is not the only kind of implicit bias training available, and we favor an alternative approach that does not depend for its success on showing that individual bias-reduction techniques are effective.

The sort of training we favor remains currently untested. Our preferred form of implicit bias training uses implicit bias as a way of motivating engagement with the ways that structures, institutions, and cultures can perpetuate marginalization and underrepresentation. We sketch a variety of institutional, cultural, and procedural changes that might be implemented but—most importantly—we invite participants to begin the process of reflecting on how their workplaces are organized and run. Crucially, we do not consider training to be completed at the end of one of these sessions. Instead, we urge participants to follow up with us to discuss ways that their specific processes can be improved; and they in fact do follow up with us at quite high rates. We know that this has very often led to substantial procedural changes in these workplaces. This sort of training has *not yet* been empirically tested, though we are in the early stages of a real-world study of this approach.

In short, there is very good empirical reason to reject certain forms of implicit bias training. But there is very good reason to continue to implement and explore other forms that look better positioned to bring about substantial changes in workplace practice and culture.

4.5 FAILURE TO ADDRESS INTERSECTIONALITY

Finally, an important concern is that the empirical research program on implicit bias has wholly failed to grapple with the realities of intersectional oppression. Indirect measures such as the IAT have focused on one dimension of identity—gender, or race, say—and have failed to recognize the distinctive forms of oppression that target the intersections of gender and race (and other aspects of social identity). Moreover, this failure is not simply one of omission; it also, it is argued, perpetuates oppression in producing "inaccurate understandings of racial and gender discrimination" (Goff and Kahn 2013), ones that serve to exclude women of color.

For example, Goff and Kahn point out that the stereotypes that are at issue in race/stereotype IATs (those to do with violence, aggression, criminality, as well as athleticism) are those typically associated with black men rather than black women (2013, 372). They note that where researchers are aware that racist stereotypes are gendered, "to the degree they must choose a gendered target, the tendency is to choose non-White men rather than non-White women" (ibid.). Moreover, there is evidence that 'prototypical' victims of gendered discrimination are white women (Goff and Kahn 2013, 376, confirming the claims in Crenshaw 1989, 1991). This suggests that when gender stereotypes are evaluated with indirect measures, what is activated in participants' cognitions are stereotypes about white women, rather than women of color. This concern is compounded by testimonial reports (see, e.g., Williams 2014) that indicate that experiences of subtle bias reported by white women and women of color are qualitatively different. This means that the research program on implicit bias is skewed toward understanding gender bias in terms of biases white women face, and racial biases in terms of biases black men face. The specific biases or forms of discrimination that black women face are largely ignored. Similarly, the intersection of other forms of oppression—class, disability, race, age—will not be accessed by indirect measures such as the IAT. This means that the field of research is particularly poorly placed to offer findings or interventions that are relevant to individuals facing oppression along multiple dimensions.

However, while this critique is certainly apt for the vast majority of the research on implicit bias, the failure of the research program to address intersectional oppression is a contingent rather than necessary feature. There is no reason that indirect measures could not be developed to address more than one dimension of oppression, or access the particular stereotypes that individuals in different social positions face. Indeed some recent work has started to examine more closely the intersection of different kinds of oppression, and the distinctive biases and stereotypes that are activated by intersectional identities. For example a few studies have started to use indirect measures to understand people's prejudicial perceptions with respect to the relationship between race and class (Brown-Iannuzzi et al. 2017; Lei and Bodehnausen 2017). Wilson and Remedios (2017) examined the biases that operate in the context of race and sexuality, finding that in some specific contexts, black gay men are stereotyped very differently from black straight men. Perhaps unsurprisingly, the biases did not seem to operate additively, but rather interacted in a more complex way, whereby associations with black gay men's sexuality (warmth) tempered negative stereotypes associated with race (aggression). As far as we know, however, just one study has focused on indirect measures looking at the intersection of race and gender, and specifically whether stereotypes about danger and criminality are associated to the same degree with black women as has been found with black men (Thiem et al. 2019).

The take-home message, then, is not that there is anything inherent in research on implicit bias that means the research tools are inapt for addressing intersecting oppressions—but that so far, they have largely failed to do so. The conclusions we can draw from existing studies on gender biases are most likely partial—about biases that white women face. The conclusions we can draw from existing studies on racial biases are likewise most likely about the biases that black men face. Overgeneralizations on the basis of these studies should be avoided, and we should note—and remedy—the paucity of research on implicit biases about black women, women from other racially stereotyped groups, and individuals marginalized due to other intersecting forms of oppression. Implicit biases are probably more complex than current research has supposed—tracking multiple intersecting dimensions of oppression, rather than just one dimension of identity (race, or gender, or …). The challenge, then, is to develop this research in ways that do not homogenize social groups and ignore importantly different ways in which individuals may be stereotyped. This does not give us a reason to reject research on implicit biases—but rather to do more of it, better.

4.5 SUMMARY

We have considered four pressing challenges to the research program on implicit bias: about the reliability of the measures; about the predictive validity of implicit biases; about the inefficacy of bias change; and the failure to adequately address intersectional oppression. In engaging with these challenges, we get a better sense of how to understand the measures, about why we might care about implicit biases, as well as how we might best focus our efforts in understanding and addressing them. In short, we argued that there is no reason to suppose the tests unreliable; that there are reasons to care about cognitions that have low predictive validity; that it is unsurprising that behavioral change is not mediated by changes in bias, given what is known elsewhere about how to bring about changes in behaviors. Moreover, there is reason to hope that future research can do better at addressing intersecting oppressions.

All this is to defend the research program on implicit bias from challenges, and establish that, in our view, it is legitimate to appeal to considerations about implicit bias to motivate reforms. Indeed, in the next section we set out some advantages of doing so.

5. WHY TALK ABOUT IMPLICIT BIASES AT ALL?

Since one could motivate many of the reforms without appeal to implicit bias (as outlined in section 3), and since one might have to deal with critiques such as those outlined in the previous sections, is there any reason to discuss implicit bias in this endeavor at all? In this section, we briefly point to reasons for which it may be efficacious to appeal to the research on implicit bias.[55]

First, appealing to considerations about implicit bias can be an effective tool for motivating people to consider reforms. One reason for this, perhaps, is that it makes clear that individuals—notwithstanding their good intentions—can be implicated in discrimination. Recognizing this, and that one may fail to live up to one's egalitarian principles, can be a significant motivator. These concerns can be particularly gripping for philosophers, since we like to think of ourselves as objective. Being presented with

[55] These reasons are elaborated and discussed in the context of other concerns elsewhere (Saul forthcoming).

research that shows that we are, most likely, failing to live up to our self-conceptions is an important corrective to our lack of self-knowledge, and can be particularly motivating.

Second, proper understandings of implicit bias enable us to connect individual cognition with broader issues to do with social structures and social injustices (Saul forthcoming). It is important both to recognize implicit bias as caused, in part, by wider social structures; and to see implicit bias as playing a role in sustaining those social structures. Focusing on the relationship between individual cognition, social structures, and actions individuals can take to try to change those social structures, can be facilitated by discussions of implicit bias—if done well (ibid.).

Third, discussing implicit bias (in the right way) can be particularly useful for motivating robust institutional change. While it is true that many of the changes advocated in section 3 can be multiply justified, the importance of adopting these strategies is most forcefully seen, perhaps, when one considers the possibility of implicit bias affecting outcomes. Consider the recommendations for changing discussions outlined above. Putting in place rules that encourage constructive and inclusive discussion of course seems like a good idea from the points of view of simply being nice and getting good philosophy done. But the extent to which the recommended *rules* are needed—rather than just commitments to try and be nice, constructive, and inclusive—is clearer once we consider the possibility of implicit bias. The research shows us that, most likely, goodwill alone will not get us there. Rules that remove discretion do a better job of combatting implicit bias. It's great that they also show an institutional commitment to being constructive and addressing marginalization.

In sum: the reform proposals are multiply justified and don't depend for their justification on the research program on implicit bias. But in any case, the research program can be defended in light of recent skeptical challenges. So, we should continue to push for reforms, and we can motivate these efforts by appealing to, inter alia, research on implicit bias.

Acknowledgements:

The authors would like to thank, for constructive discussion and feedback, audiences at The University of Edinburgh, Bochum University, McGill University, University of Hanover, and The University of Sheffield. Thanks also to Alex Madva and Michael Brownstein for helpful discussions, and in particular to Kate Manne and Bianka Takaoka.

REFERENCES

Abraham, Charles, and Paschal Sheeran. 2005. “The Health Belief Model.” *Predicting Health Behaviour* 2: 28–80.

Amodio, D.M. and Devine, P.G., 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), p.652.

Antony, L. 2012. “Different Voices or Perfect Storm? Why Are There So Few Women in Philosophy?” *Journal of Social Philosophy* 43(3): 227–55.

Baron, S., T. Dougherty, and K. Miller. 2015. “Why Is There Female Under-Representation among Philosophy Majors?” *Ergo* 2(14).

Beebee, H. 2013. “Women and Deviance in Philosophy.” In *Women in Philosophy: What Needs to Change?* edited by K. Hutchison and F. Jenkins, pp. 51–61. Oxford: Oxford University Press.

Beebee, H., and J. Saul. 2011. “Women in Philosophy in the UK.” A Report by the British Philosophical Association and the Society for Women in Philosophy in the UK.

Beebee, H., and J. Saul. 2014. “Good Practice Guidelines for Women in Philosophy.” A Report by British Philosophical Association and Society for Women in Philosophy in the UK.

Bertrand, M., and S. Mullainathan. 2004. “Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review* 94(4): 991–1013.

Botts, T., L. K. Bright, M. Cherry, G. Mallarangeng, and Q. Spencer. 2014. “What Is the State of Blacks in Philosophy?” *Critical Philosophy of Race* 2(2): 224–42.

Bradley, C. 1984. “Sex Bias in the Evaluation of Students.” *British Journal of Social Psychology* 23(2): 147–53.

Bradley, C. 1993. “Sex Bias in Student Assessment Overlooked?” *Assessment and Evaluation in Higher Education* 18(1): 3–8.

Brown-Iannuzzi, J.L., Dotsch, R., Cooley, E. and Payne, B.K., 2017. The relationship between mental representations of welfare recipients and attitudes toward welfare. *Psychological science*, *28*(1), pp.92-103.

Brownstein, M. 2015. “Michael Brownstein on Williams and Ceci.” https://feministphilosophers.wordpress.com/2015/04/15/michael-brownstein-on-williams-and-ceci/#comment-139224.

Brownstein M. ms. Skepticism about Implicit Bias.

Brownstein, M., A. Madva, and B. Gawronski. ms. Understanding Implicit Bias: Putting the Criticism into Perspective.

Budden, A., et al. 2008. “Double-Blind Review Favours Increased Representation of Female Authors.” *Trends in Ecology and Evolution* 23(1): 4–6.

Calhoun, C. 2009. “The Undergraduate Pipeline Problem.” *Hypatia* 24(2): 216–23.

Carlana, Michela. 2018. “Implicit Stereotypes: Evidence from Teachers’ Gender Bias.” HKS Working Paper NO. RWP18-034. Available at SSRN: https://ssrn.com/abstract=3274717 or http://dx.doi.org/10.2139/ssrn.3274717

Carnes, M., et al. 2015. “Effect of an Intervention to Break the Gender Bias Habit for Faculty at One Institution: A Cluster Randomized, Controlled Trial.” *Academic Medicine: Journal of the Association of American Medical Colleges* 90(2): 221–30.

Crenshaw, Kimberle (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *The University of Chicago Legal Forum* 140:139-167.

Crenshaw, Kimberle Williams, (1991). ‘Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color’. *Stanford Law Review* 43 (6):1241-99.

Danziger, S., J. Levav, and L. Avnaim-Pesso. 2011. “Extraneous Factors in Judicial Decisions.” *Proceedings of the National Academy of Sciences* 108(17): 6889–92.

Dasgupta, N. and Asgari, S., 2004. Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of experimental social psychology*, *40*(5), pp.642-658.

Dasgupta, N. 2011. "Ingroup Experts and Peers as Social Vaccines Who Inoculate the Self-Concept: The Stereotype Inoculation Model." *Psychological Inquiry* 22(4): 231–46.

Devine, Patricia, et al. 2012. "Long-Term Reduction in Implicit Race Bias: A Prejudice Habit-Breaking Intervention." *Journal of Experimental Social Psychology* 48: 1267–78.

Di Bella, Laura, Eleanor Miles, and Jennifer Saul. 2016. "Philosophers Explicitly Associate Philosophy with Maleness." In *Implicit Bias and Philosophy Volume I: Metaphysics and Epistemology*, edited by Michael Brownstein and Jennifer Saul, 283–308. Oxford: Oxford University Press.

Dotson, Kristie. 2018. "On Intellectual Diversity and Differences That May not Make a Difference." *Ethics and Education* 13(1): 123–40.

Doucet, M., and G. Beaulac. 2013. "Report on the CPA Equity Survey: Historical Trends." Canadian Philosophical Association.

Dougherty, T., S. Baron, and K. Miller. 2015. "Why Do Female Students Leave Philosophy? The Story from Sydney." *Hypatia* 30(2): 467–74.

Duguid, Michelle M., and Melissa C. Thomas-Hunt. 2015. "Condoning Stereotyping? How Awareness of Stereotyping Prevalence Impacts Expression of Stereotypes." *Journal of Applied Psychology* 100(2): 343.

Duncan, Birt L. 1976. "Differential Social Perception and Attribution of Intergroup Violence: Testing the Lower Limits of Stereotyping of Blacks." *Journal of Personality and Social Psychology* 34(4): 590.

Dutt, K., et al. 2016. "Gender Differences in Recommendation Letters for Postdoctoral Fellowships in Geoscience." *Nature Geoscience* 9(11): 805–808.

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

Enqvist, L., and J. G. Frommen. 2008. “Double-Blind Peer Review and Gender Publication Bias.” *Animal Behaviour* 76(3): e1–e2.

Forscher, Patrick S., et al. 2017a. “A Meta-analysis of Change in Implicit Bias.” *Psychological Bulletin*.

Forscher, Patrick S., et al. 2017b. “Breaking the Prejudice Habit: Mechanisms, Timecourse, and Longevity.” *Journal of Experimental Social Psychology* 72: 133–46.

Friedman, M. 2013. “Women in Philosophy: Why Should We Care?” In *Women in Philosophy: What Needs to Change?* edited by K. Hutchison and F. Jenkins, 21–38. Oxford: Oxford University Press.

Gawronski, Bertram. In press. “Six Lessons for a Cogent Science of Implicit Bias and Its Criticism.” *Perspectives on Psychological Science*.

Glöckner, A. 2016. “The Irrational Hungry Judge Effect Revisited: Simulations Reveal That the Magnitude of the Effect Is Overestimated.” *Judgment and Decision Making* 11(6): 601–10.

Goff, Phillip Atiba, and Kimberly Barsamian Kahn. 2013. “How Psychological Science Impedes Intersectional Thinking.” *Du Bois Review: Social Science Research on Race* 10(2): 365–84.

Green, A.R., Carney, D.R., Pallin, D.J., Ngo, L.H., Raymond, K.L., Iezzoni, L.I. and Banaji, M.R., 2007. Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of general internal medicine*, *22*(9), pp.1231-1238.

Greenwald, A.G., McGhee, D.E. and Schwartz, J.L., 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), p.1464.

Greenwald, A.G., Banaji, M.R. and Nosek, B.A., 2015. Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of personality and social psychology*, *108*(4), pp.553-561.

Haslanger, S. 2008. "Changing the Ideology and Culture of Philosophy: Not by Reason (Alone)." *Hypatia* 23(2): 210–23.

Haslanger, S. 2009. "Preliminary Report of the Survey on Publishing in Philosophy." APA Committee on the Status of Women in Philosophy.

Healey, K. 2015. "Gender and Citation in Four General Interest Philosophy Journals 1993–2013." https://kieranhealy.org/blog/archives/2015/02/25/gender-and-citation- in-four-general-interest-philosophy-journals-1993-2013/.

Hermanson, S. 2017. "Implicit Bias, Stereotype Threat, and Political Correctness in Philosophy." *Philosophies* 2(12): 1–17.

Hinton, D.P. and Higson, H., 2017. A large-scale examination of the effectiveness of anonymous marking in reducing group performance differences in higher education assessment. *PloS one*, *12*(8), p.e0182711.

Holroyd, J., and K. Puddifoot. ms. Bias, Testimony and Epistemic Injustice.

Jenkins, F. 2013. "Singing the Post-Discrimination Blues: Notes for a Critique of Academic Meritocracy." In *Women in Philosophy: What Needs to Change?* edited by K. Hutchison and F. Jenkins, 81–102. Oxford: Oxford University Press.

Jost, John T. 2018. “The IAT Is Dead, Long Live the IAT: Context-Sensitive Measures of Implicit Attitudes Are Indispensable to Social and Political Psychology.” *Current Directions in Psychological Science* 28 (1), 10-19.

Knobloch-Westerwick, S., Glynn, C.J. and Huge, M., 2013. The Matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication*, *35*(5), pp.603-625.

Lee, C., et al. 2013. “Bias in Peer Review.” *Journal of the American Society for Information Science and Technology* 64(1): 2–17.

Lei, R.F. and Bodenhausen, G.V., 2017. Racial assumptions color the mental representation of social class. *Frontiers in psychology*, *8*, p.519.

Leslie, Sarah-Jane, et al. 2015. “Expectations of Brilliance Underlie Gender Distributions across Academic Disciplines.” *Science* 347(6219): 262–65.

McClean, Elizabeth J., et al. 2018. “The Social Consequences of Voice: An Examination of Voice Type and Gender on Status and Subsequent Leader Emergence.” *Academy of Management Journal* 61(5): 1869–91.

Madera, J., M. Hebl, and R. Martin. 2009. “Gender and Letters of Recommendation for Academia: Agentic and Communal Differences.” *Journal of Applied Psychology* 94(6): 1591–99.

Madva, A. 2017. “Biased against De-Biasing.” *Ergo* 4: 6. http://dx.doi.org/10.3998/ergo.12405314.0004.006.

Maliniak, D., et al. 2013, October. “The Gender Citation Gap in International Relations.” *International Organization* 67(4): 889–922.

Merton, R. 1968. “The Matthew Effect in Science.” *Science* 159(3810): 56–63.

Mills, C. 2005. “‘Ideal Theory’ as Ideology.” *Hypatia* 20(3): 165–84.

Moulton, J. 1983. “A Paradigm of Philosophy: The Adversary Method.” In *Discovering Reality*, edited by S. Harding and M. Hintikka, 149–64. Dordecht: Riedel.

Moss-Racusin, C., J. Dovidio, V. Brescoll, M. Graham, and J. Handelsman. 2012. “Science Faculty’s Subtle Gender Biases Favor Male Students.” *PNAS* 109(41): 16474–79.

Norlock, K., 2011, February. Women in the Profession: A Report to the CSW. In *American Philosophical Association*.

Olson, M.A. and Fazio, R.H., 2006. Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*(4), pp.421-433.

O’Neill, E. 1997. “Disappearing Ink: Early Modern Women Philosophers and Their Fate in History.” In *Philosophy in a Feminist Voice*, edited by J. Kourany, 17–62. Princeton, NJ: Princeton University Press.

Oswald, F.L., Mitchell, G., Blanton, H., Jaccard, J. and Tetlock, P.E., 2013. Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of personality and social psychology*, *105*(2), p.171.

Paxton, M., C. Figdor, and V. Tiberius. 2012. “Quantifying the Gender Gap: An Empirical Study of the Underrepresentation of Women in Philosophy.” *Hypatia* 27(4): 949–57.

Payne, B. Keith, Heidi A. Vuletich, and Kristjen B. Lundberg. 2017. “The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice.” *Psychological Inquiry* 28(4): 233–48.

Payne, B. Keith. 2001. “Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon.” *Journal of Personality and Social Psychology* 81(2): 181.

Peters, Douglas P., and Stephen J. Ceci. 1982. “Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again.” *Behavioral and Brain Sciences* 5: 187–255.

Pitt, Edd, and Naomi Winstone. 2018. “The Impact of Anonymous Marking on Students' Perceptions of Fairness, Feedback and Relationships with Lecturers.” *Assessment & Evaluation in Higher Education*: 43 (7): 1183–1193.

Prestwich, A., Perugini, M. and Hurling, R., 2009. Can the effects of implementation intentions on exercise be enhanced using text messages?. *Psychology and Health*, *24*(6), pp.677-687.

Rooney, P. 2009. “Philosophy, Adversarial Argumentation, and Embattled Reason.” *Informal Logic* 30(3): 203–34.

Rooth, Dan-Olof. 2010. “Automatic Associations and Discrimination in Hiring: Real World Evidence.” *Labour Economics* 17(3): 523–34.

Saul, J. Forthcoming. “(How) Should We Tell Implicit Bias Stories?” *Disputatio*.

Saul, J. 2013. “Implicit Bias, Stereotype Threat and Women in Philosophy.” In *Women in Philosophy: What Needs to Change?* edited by K. Hutchison and F. Jenkins, 39–60. Oxford: Oxford University Press.

Schwitzgebel, E. 2010. “On Being Good at Seeming Smart.” http://schwitzsplinters.blogspot.com/2010/03/on-being-good-at-seeming-smart.html.

Steinpreis, R., K. Anders, and D. Ritzke. 1999. "The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study." *Sex Roles* 41(7/8): 509–28.

Stewart, A., and V. Valian. 2018. *An Inclusive Academy: Achieving Diversity and Excellence*. Cambridge, MA: MIT Press.

Stewart, B.D. and Payne, B.K., 2008. Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, *34*(10), pp.1332-1345.

Thiem, Kelsey C., et al. 2019. "Are Black Women and Girls Associated With Danger? Implicit Racial Bias at the Intersection of Target Age and Gender." *Personality and Social Psychology Bulletin*. DOI: 0146167219829182.

Thompson, M., T. Adleberg, S. Sims, and E. Nahmias. 2016. "Why Do Women Leave Philosophy? Surveying Students at the Introductory Level." *Philosophers' Imprint* 16(6).

Uhlmann, E., and G. Cohen. 2007. "'I think it, therefore it's true': Effects of Self-Perceived Objectivity on Hiring Discrimination." *Organizational Behavior and Human Decision Processes* 104 (2): 207–23.

Webb, T., O'Hara, B., and R. Freckleton. 2008. "Does Double-Blind Peer Review Benefit Female Authors?" *Trends in Ecology and Evolution* 23(7): 351–53.

Wenneras, C., and A. Wold. 1997. "Nepotism and Sexism in Peer Review." *Nature* 387: 341–43.

Whittaker, R. 2008. "Journal Review and Gender Equality." *Trends in Ecology and Evolution* 23(9): 478–79.

Williams, Joan C. 2014. “Double Jeopardy? An Empirical Study with Implications for the Debates over Implicit Bias and Intersectionality.” *Harvard Journal of Law & Gender* 37: 185.

Williams, W., and P. Ceci. 2015. “National Hiring Experiments Reveal 2:1 Preference for Women on STEM Tenure Track.” *PNAS* 112(17): 5360–65.

Wilson, J.P., Remedios, J.D. and Rule, N.O., 2017. Interactive effects of obvious and ambiguous social categories on perceptions of leadership: When double-minority status may be beneficial. *Personality and Social Psychology Bulletin*, *43*(6), pp.888-900.