JENNIFER SAUL

Scepticism and Implicit Bias

Who Is Jennifer Saul?

Jennifer Saul is an American-born philosopher who worked for many years at the University of Sheffield in the UK and has recently moved to become Waterloo Chair in Social and Political Philosophy of Language at the University of Waterloo in Ontario, Canada. She is known for her work in both philosophy of language and feminist philosophy. Saul was born in Ohio to a progressive family of academics. She recalls attending her first feminist consciousness-raising meeting at four years old, accompanied by her grandmother, a mathematician who was active in the 1970s feminist movement.* She earned her MA and PhD at Princeton, specializing in analytic philosophy of language. As Saul notes, she hadn't yet entertained the possibility of feminist philosophy during her student years, believing that feminism was "too obviously correct to be something one could do philosophy about."† Ultimately, Saul came to recognize that social justice is connected in important ways to the very aspects of language that were already the focus of her work.

Saul's publications include *Feminism: Issues & Arguments* (2003), *Substitution, Simple Sentences and Intuitions* (2007), and *Lying, Misleading and What Is Said: An Exploration in Philosophy of Language and in Ethics* (2012). Saul co-founded two influential blogs: "Feminist Philosophers" and "What Is It Like to Be a Woman in Philosophy?"‡ She has been a vocal advocate for remedying the race and gender imbalance in academic philosophy, and in 2011 was awarded the Distinguished Woman Philosopher Award by the Society for Women in Philosophy.

What Is the Structure of This Reading?

Saul begins this paper by reviewing the scientific literature surrounding the phenomenon of implicit bias, illustrating a number of ways in which implicit biases appear to disadvantage some people in hiring processes, academic work, and other social interactions. She then argues that the doubt generated by the discovery of implicit bias entails a special sort of skepticism. While traditional philosophical skepticism is often seen as a form of "armchair philosophy," with little practical significance in our lives, the skepticism arising from bias-related doubt is far more troublesome, as it demands both attention and action. According to Saul, we ought to be more "unsettled" by implicit bias than by more traditional skeptical concerns such as the possibility that we are brains in vats. While we may have difficulty proving definitively that we are not brains in vats, in the case of bias-related doubt we actually have strong positive evidence from scientific studies indicating that we are frequently making errors. Moreover, if we are making bias-related errors we really should take action to change this given the harmful social consequences, whereas if we are brains in vats it's not clear that this would actually require us to change our actions.

In the final sections of this paper, Saul asks what we should do about these new concerns. We might attempt to use "counter-intuitive mechanical techniques" to reduce the effects of implicit bias, but Saul notes that the effectiveness of such techniques may be limited, as they typically target very specific associations and behaviors. The only complete solution, according to Saul, is to eradicate all types of prejudice and stereotype, which would result in nothing less than a "sweeping and radical transformation of our social world."

---

* Interview with Jennifer Saul, February 17, 2007, on the "What Is It Like to Be a Philosopher?" blog.

† Ibid.

‡ The title of the latter is an allusion to Thomas Nagel's "What Is It Like to Be a Bat?" (included in the Philosophy of Mind chapter of this volume).

Scepticism and Implicit Bias[*]

The goal of this paper is to explore the idea that what we know about implicit bias gives rise to something *akin to* a new form of scepticism. I am not wedded to the idea that the phenomenon I am pointing to should be called 'scepticism', but I am convinced that it is illuminating to examine the ways in which it does and does not resemble philosophical scepticism. I will call what I am discussing 'bias-related doubt'.

In some ways, bias-related doubt is stronger than traditional forms of scepticism, while in others it is weaker. In brief: I will be arguing that what we know about implicit biases shows us that we have very good reason to believe that we cannot properly trust our knowledge-seeking faculties. This does not mean that we might be mistaken *about everything*, or even everything in the external world (so it is weaker than traditional scepticism). But it does mean that we have *good reason* to believe that we are mistaken about a great deal (so it is stronger than traditional forms of scepticism). A further way in which bias-related doubt is stronger than traditional scepticism: this is doubt that demands action. With traditional scepticism, we feel perfectly fine about setting aside the doubts we have felt when we leave the philosophy seminar room. But with bias-related doubt, we don't feel fine about this at all. We feel a need to *do something* to improve our epistemic situation. Fortunately, though, it turns out that there is much we can do. However, much of what needs to be done cannot be done on a purely individual basis. So although scepticism has sometimes been treated by feminists as a paradigmatic case of the excesses of individualist philosophy,[1] this form of scepticism cannot be fully responded to individualistically.

1 Implicit Biases

There is a vast and still-growing literature on implicit bias, which I'll only be dipping into here. Very broadly speaking, these are largely unconscious tendencies to automatically associate concepts with one another.[2] Put like this, they don't sound very interesting or worrying. But the ones on which attention by philosophers has focused are both very interesting and very worrying. These are unconscious, automatic tendencies to associate certain traits with members of particular social groups, in ways that lead to some very disturbing errors: we tend to judge members of stigmatized groups more negatively, in a whole host of ways. Rather than attempt a general overview, I will give examples of the sorts of errors that will be our concern here.

CURRICULUM VITAE

CV studies take a common, and beautifully simple form. The experimenters ask subjects to rate what is in fact the same CV, varying whatever trait they want to study by (usually) varying the name at the top of it. When they do this, they find that the same CV is considered much better when it has a typically white rather than typically black name, a typically Swedish rather than typically Arab name, a typically male rather than typically female name, and so on. The right name makes the reader rate one as more likely to be interviewed, more likely to be hired, likely to be offered more money, and a better prospect for mentoring. These judgments are very clearly being affected by something that *should* be irrelevant—the social category of the person whose CV is being read. Moreover, the person making these mistaken judgments is surely unaware of the role that social category is playing in the formation of their views of the candidates. Significantly, the most recent of these studies (Moss-Racusin 2012), on the evaluation of women's CVs, showed that women were just as likely to make these problematic judgments as men. It also showed that these problems are not confined to an older generation: the tendencies were equally strong in all age groups.[3]

PRESTIGE BIAS

In a now-classic study, psychologists Peters and Ceci[4] (1982) sent previously published papers to the top

[*] Jennifer Saul, "Scepticism and Implicit Bias," *Disputatio* 5, 37 (2013): 243–63.

psychology journals that had published them, but with false names and non-prestigious affiliations. Only 8% detected that the papers had already been submitted, and 89% were rejected, citing serious methodological errors (and not the one they should have cited—plagiarism). This makes it clear that institutional affiliation has a dramatic effect on the judgments made by reviewers (either positively, negatively, or both). These are experts in their field, making judgments about their area of expertise—psychological methodology—and yet they are making dramatically different judgments depending on the social group to which authors belong (member of prestigious vs non-prestigious psychology department).

PERCEPTION

Studies of so-called 'shooter bias' show us that implicit bias can even influence perception. In these studies, it has been shown that the very same ambiguous object is far more likely to be perceived as a gun when held by a young black man and something innocent (like a phone) when held like by a young white man.[4] (The same effect has been shown with men who appear Muslim versus men who appear non-Muslim (Unkelbach et al. 2008). In some of these experiments, the subjects' task is to shoot in a video game if and only if they see an image of a person carrying a gun. Subjects' 'shooting' is just as you'd expect given their perceptions. These show that implicit bias is getting to us even before we get to the point of reflecting upon the world—it affects our very perceptions of that world, again in worrying ways.[5]

MORAL AND POLITICAL CONSEQUENCES

Now let's explore some consequences of this. First, there are some obvious morally and politically significant consequences. We are very likely to make inaccurate judgments about who is the best candidate for a job, if some of the top candidates are known to be from stigmatised groups. We are very likely to mark inaccurately, if social group membership is known to us and the group we are marking is not socially homogeneous. We are very likely to make inaccurate judgments about which papers deserve to be published, if social group membership is known to us. We may both over-rate members of some groups and under-rate others. Worse yet, we are misperceiving harmless objects as dangerous, and potentially acting on this in truly appalling ways. All of this *should* be tremendously disturbing to us. It means that we are being dramatically *unfair* in our judgments, even though we are doing so unintentionally. We are treating members of stigmatised groups badly, even if we desperately desire to treat them well. Moreover, what we are doing will help to ensure that this unfair treatment is continued: the results of these decisions will help to maintain the stereotypes that currently exist, which cause members of stigmatised groups to be treated unfairly. 'Vicious circle' seems a particularly apt phrase to describe the situation.

EPISTEMOLOGICAL CONSEQUENCES

But I want to focus now on some epistemological aspects of this situation. First, some relatively obvious ones, starting from those within philosophy. The unfairness described above means that there are almost certain to be some excellent students receiving lower marks and less encouragement than they should; some excellent philosophers not getting the jobs they should get; and where anonymous refereeing and editing is not practised, there is some excellent work not being published. Philosophy as a field is the worse for this: it is not as good as it could, or should, be. (For more on this, see Beebee and Saul 2011, Saul forthcoming.) Obviously, much the same will go on in other areas of academia, especially those that are as male-dominated as philosophy. Outside philosophy, there are similar effects, as the testimony of members of stigmatised groups is taken less seriously than it ought to be (Fricker 2007). Their views are less respected, and they are given less of an opportunity to participate fully in discussions and decision-making. As Chris Hookway (2010) has noted, a particular problem may lie in their *questions* not being taken seriously.

Now, some less obvious epistemological aspects of the situation, again focussing on philosophy. When we misjudge a paper's quality, we're making a mistake about the quality of an argument.[6] Moreover, our evaluation of that argument is being influenced by factors totally irrelevant to its quality: it's being influenced by our knowledge of the social group of its author. Worse yet, this influence operates below the level of

consciousness—it's unavailable to inspection and rational evaluation. This means we may be accepting arguments we should not accept and rejecting arguments we should not reject. Many of our philosophical beliefs—those beliefs we take to have been arrived at through the most careful exercise of reason—are likely to be wrong.[7]

It is important to see that this is not *just* a matter of what Miranda Fricker (2007) has called testimonial injustice. Fricker argues that the social group to which a person belongs will often have a dramatic effect on our willingness to treat them as a credible source of knowledge. We will be less likely to accept the testimony of those from stigmatised groups. One thing implicit bias adds to this picture is just a matter of scale: research shows these problems to be far more widespread than would otherwise be apparent. But another, even more important addition, is that implicit bias doesn't just affect our judgments of people's *credibility* when deciding whether to accept their testimony or not. Mistaken as our credibility judgments are, at least we know that these are judgments about who to take seriously. We recognise that we are making judgments about people, and this is what we mean to be doing. The research on implicit bias shows us that we are actually being affected by biases about social groups *when we think we are evaluating evidence or methodology*. When considering testimony, it makes sense that we need to make judgments about how credible an individual is. But when psychologists assess the methodology of a study—or when philosophers assess the quality of an argument—they shouldn't be looking at the credibility of an individual at all. They should be looking just at the study, or the argument. And yet when implicit bias is at work, we are likely to be affected by the social group of the person presenting evidence or an argument even when we are trying to evaluate that evidence or argument itself. Implicit bias is not just affecting who we trust—it's affecting us when we think we're making judgments that have nothing to do with trust. It's leading us into errors based on social category membership when we think we're making judgments of scientific or argumentative merit.

But why should that unsettle us? We know already that most of what is currently accepted as science is likely to be proven false within centuries, and possibly decades. But notice: my claim is not that we're likely to be accepting some falsehoods, or even a lot of falsehoods. That's not unsettling. My claim is that we're likely to be *making errors*. Moreover, we're likely to be making errors of a very specific sort. It's *not* that we're likely to get some really difficult technical bits wrong, or that we're likely to get things wrong if we're really exhausted, or drunk. It's that we're likely to let the social identity of the person making an argument affect our evaluation of that argument. It is part of our self-understanding as rational enquirers that we will make certain sorts of mistakes. But not this sort of mistake. These mistakes are ones in which something that we actively think *should not* affect us does.

Worse yet, our errors are not confined to the professional arena, or to what we take to be carefully thought-out judgments about the quality of arguments that we encounter. The studies of shooter bias show us that as humans in the world, we are making errors in *perception* due to implicit bias. The very data from which we begin in thinking about the world—our perceptions—cannot be relied upon to be free of bias. Once more, this is clearly well beyond the worries raised by testimonial injustice.

The best way to see why these mistakes are—and should be—so unsettling to us as enquirers is to compare the situation of one who learns about implicit biases to the situations of people considering various sorts of sceptical scenarios.

2 Comparison to Sceptical Scenarios

2.1 COMPARISON TO TRADITIONAL SCEPTICISM

In a traditional sceptical scenario, we are confronted with a possibility that we can't rule out—that we're brains in vats,* or that tomorrow gravity might not

* This is an updated version of René Descartes's evil demon thought experiment; see Gilbert Harman, *Thought* (Princeton University Press, 1973), 5.

work any more.* Considering this scenario is meant to make us worry that we don't know (many of) the things that we take ourselves to know, or that we are unjustified in having (many of) the beliefs that we do. And a standard response is that these worries should not grip us, because we have no reason at all to suppose that these possibilities obtain. Doubt induced by implicit bias is unlike this: we have *very good reason* to suppose that we are systematically making errors caused by our unconscious biases related to social categories. In this way, then, the doubt provoked by implicit bias is stronger than that caused by considering sceptical arguments.

But, one might think, it's not really all that troubling. The doubt caused by implicit bias, surely, is a localized one. It seems, at first, to be like the sort of doubt we experience when we discover how poor we are at probabilistic reasoning. We have extremely good reason to think we're making errors when we make judgments of likelihood. But this sort of doubt doesn't trouble us all that much because we know exactly when we should worry and what we should do about it: if we find ourselves estimating likelihood, we should mistrust our instincts and either follow mechanical procedures we've learned or consult an expert (if not in person, then on the internet). This kind of worry is one that everyone can accept without feeling drawn into anything like scepticism. And it may seem at first that bias-related doubt is like this.

The problem starts to become vivid when we ask ourselves *when* we should be worried about implicit bias influencing our judgments. The answer is that we should be worried about it whenever we consider a claim, an argument, a suggestion, a question, etc. from a person whose apparent social group we're in a position to recognize. Whenever that's the case, there will be room for our unconscious biases to perniciously affect us. Most discussed in the literature so far (see Fricker 2007), we might make a mistaken judgment of credibility when assessing testimony. But we also might fail to listen properly to a contribution; fail to carefully consider a question; judge an argument to be less compelling or original than it is; think the evidence presented is worse than it is. And, importantly, we can be adversely affected in a positive direction as well. When assessing a contribution from someone who our biases favour, we may grant more credibility than their testimony deserves; we may think their arguments are better than they are, perhaps failing to notice flaws that we would have noticed if the arguments were presented by someone else; we may take their evidence to be better than it is, and so on.

And *this* is going to happen a great deal. It happens whenever we are dealing with the social world in a non-anonymised† manner. Since the world is only rarely anonymised for us, this will happen nearly all the time. Much of our knowledge comes from testimony, or from arguments or evidence that we are presented with. Those testifying, or presenting the arguments or evidence, are usually people. And people are generally (though not always) perceived by us as members of social groups. Moreover, much of the knowledge we already have has come to us in this way. Our acceptance or rejection of testimony, arguments, evidence and the like has shaped the worldviews we have now. And this acceptance or rejection was, we can be fairly certain, distorted by the perceived social groups[8] of those presenting the testimony, arguments or evidence. Worse yet, we cannot even go back and attempt to consider or correct errors that we might have made—we are very unlikely to remember the sources of these beliefs of ours.

...

[Chris] Hookway writes that there are three key features to 'an interesting sceptical challenge'. (1990: 164)

1. It must make reference to 'part of our practice of obtaining information about our surroundings which we find natural, which it does not ordinarily occur to us to challenge.'
2. '[I]t must have a certain generality: challenges to the reliability of particular thermometers may lead us to lose confidence in that particular instrument; they do not lead us to lose confidence in ourselves as inquirers.'

---

* This is an allusion to Hume's problem of induction (see the Hume selection in the Philosophy of Science section of this volume).

† Anonymized material has all information about who it came from removed.

3. '[I]t must intimate that the feature of our practice which it draws attention to *could not* be defended.'

It seems to me that bias-related doubt easily meets each of these criteria. The practices called into question are ones that we normally don't think to question: our 'instinctive' sense that someone is credible, that a reason is convincing, or that an argument is compelling. There is definitely generality—this isn't like challenges just to probabilistic reasoning, which Hookway rightly flags as not that worrying because those challenges are very contained. Instead, it's challenges to the ordinary ways that we assess reasons, arguments evidence and testimony. Finally, the feature it calls attention to—our judgments are illicitly influenced by irrelevant matters in a way that frequently leads to injustice—is deeply indefensible.

What the literature on implicit bias shows us is that we *really should not* trust ourselves as inquirers. As Hookway argues (2003: 200), 'we can persevere with our inquiries only if we are confident that ... our reflection will take appropriate routes'. But we have now discovered that our reflection takes wholly inappropriate routes: we are not only failing to assess claims or arguments by methods that we endorse but we are instead assessing them by methods that we actively oppose. As he notes, only a part of the process of deliberation is conscious, and we need to be able to trust the habits of thought that underpin the unconscious bits (Hookway 1990: 11). We need to trust not just that they will guide us to truth but that they are based in values that we consider our own. Hookway raises the values concern when discussing an obsessive who is unable to stop repeatedly rehearsing doubts that he does not fully endorse, but the concern arises even more strongly in the case of biases against members of stigmatized groups. The literature on implicit bias shows us not just that our habits can't be relied on to lead us to truth, but also that—insofar as they can be described as based in values at all—they are likely to be based in values that we (most of us, anyway) find repugnant. It is difficult to see how we could ever properly trust these again once we have reflected on implicit bias. And, Hookway (2000: Chapter 10) argues, self-trust is a necessary condition of responsible inquiry.

2.2 COMPARISON TO LIVE SCEPTICAL SCENARIOS

Bryan Frances's work on 'live sceptical scenarios' (Frances 2005), provides another instructive comparison. Frances characterizes traditional sceptical arguments as relying on the fact that certain hypotheses cannot be ruled out. He notes that responses to these often involve pointing out that, while these hypotheses cannot be ruled out, they are nonetheless not really *live*—they are so implausible that we can't really take them seriously. His book is devoted to arguing that there are sceptical hypotheses that are not like this. In his live sceptical scenarios, 'there are compelling scientific and philosophical reasons to think that the hypotheses are actually true'. Therefore, the traditional replies do not apply.

Now this looks quite a lot like what I have called Bias-Related Doubt. The hypotheses are ones for which there is compelling reason for thinking that they are true. But on closer inspection, it turns out that these reasons are far less compelling. The hypotheses in question are things like eliminativism* about belief and error theory about colour.† And the reasons for thinking that they are still live is that some sensible people who know a great deal endorse (or might endorse) these theories on the grounds of compelling scientific or philosophical reasons. But this falls a good deal short of what I have argued about bias-related doubt. Here the hypothesis is that we are frequently making errors that have their root in implicit bias. My claim is not just that the hypothesis is live—that sensible and knowledgeable people might endorse it on the basis of good reasons. Instead, it's that *we all have very good reason to believe that it is true.* And this is

* Eliminativism about beliefs is the view that beliefs (and, usually, other mental states such as desires, thoughts, etc.) do not in fact exist; when we talk about a person "having a belief," we are simply mistaken; beliefs, etc., are not constituted by brain states or independent mental states either. Like witchcraft, they simply do not exist.

† Error theory (about color) is the view that colors do not exist in reality, and that when we make statements about certain objects having certain colors (e.g., "The sky is blue") we are mistaken.

much stronger than the claim that a hypothesis is live. We will see that there are also differences with regard to how we should respond.

3 What Should We Do?

The scepticism created by learning about implicit bias differs dramatically from most other forms of scepticism in that it leads to the conclusion that we should change our behaviour. A striking feature of the sorts of scepticism that have tended to dominate discussion in recent times is that *even if* we became convinced by them, we would not feel the need to change anything about our behaviour: accepting that I don't know whether I'm a brain in a vat or not simply doesn't affect how I will go about living my life. Becoming a sceptic of the traditional sort doesn't lead me to decide differently about anything in the course of my every day life, or to alter my behaviour in any way.

But not all forms of scepticism are like these in their lack of impact on behaviour: Pyrrhonian scepticism* was meant to have a large and salutary impact on one's life. The convinced Pyrrhonian sceptic would learn to simply accept appearances rather than striving for belief.

> 'If he avoids "belief", the Pyrrhonist "acquiesces in appearances": he is guided by sensory appearances and by bodily needs and natural desires; he conforms to the prevailing customs and standards of his society.'[9]

Accepting appearances and conforming to prevailing customs and standards, of course, is very much *not* what a would-be responsible enquirer should feel moved to do after learning about implicit bias. For the literature on implicit bias shows that the way things appear to us is perniciously affected by biases that we are unaware of and would repudiate if we became aware of them. To put it bluntly, accepting appearances would mean acquiescing in one's reaction of fear at the sight of a black man; and acquiescing in one's greater sense of approval when looking at a CV with a man's name at the top of it. That these would not rise to the level of belief may mean that we're not committed to falsehoods. But the behaviours we would be led to would be just as troubling. As Hookway notes (1990: 18), the Pyrrhonist's 'is a very conservative outlook: the appearances he relies on are salient for him because of their conventional role.' Relying on the conventions of one's society is deeply cast into doubt by the literature on implicit bias.

The scepticism produced by implicit bias demands action. There are several reasons for this. The first reason is that the sceptical scenario is one that is troubling in a very different way from more traditional sceptical scenarios. If you actually are a brain in a vat, you're probably doing about as well with your life as you can. It's not clear that you would make different choices if you knew the scenario to hold. (And this is just as true for the live sceptical scenarios Frances considers, like those based in eliminativism or colour error theory.) But if you actually are basing lots of decisions on the social categories that people you encounter belong to, then you're clearly not doing as well as you can. You're making the wrong decisions epistemically speaking: taking an argument to be better than it is, perhaps; or wrongly discounting the view of someone you should listen to. You're also making the wrong decisions practically speaking: assigning the wrong mark to an essay, or rejecting a paper that you should accept. Finally, you're making the wrong decisions morally speaking: you are treating people unfairly; and you are basing your decisions on stereotypes that you find morally repugnant. So when the possibility is raised that you're doing this, it should not be possible to shrug it off in the way that it's perfectly reasonable to shrug off the brain in a vat possibility. Worse yet, it's not just the *possibility* that's raised: the research on implicit bias suggests that it's very likely that you're doing these things, with respect to at least some social categories.

But usually, you can't do anything at all to rule out the sceptical scenarios. And the same is true when it comes to any particular instance of the implicit bias sceptical scenario. Did I judge that woman's work to be less good than it was due to her gender? I will never

---

* Pyrrho (c. 360–c. 270 BCE) was a Greek philosopher who founded the school of Pyrrhonism, or Pyrrhonian scepticism, which was influential in the ancient Greek and Roman world. Its main tenet is that we should suspend belief concerning any proposition that is not completely evident.

know, because I won't get the opportunity to assess it without knowledge of her gender. And the same is true for certain more general versions: have I based much of what I think I know on epistemically irrelevant factors like social categories? I'm not going to be able to find out. So is there *anything* one can do? Not for past cases like these. However, I can act so as to reduce the likelihood of this happening in future instances.

Importantly, though, some of the most obvious things to do just don't work. Getting a woman to judge another woman's work is a poor check against bias, since both men and women are likely to hold biases causing them to negatively judge women's work (recall Moss-Racusin's 2012 CV study). Trying hard to be unprejudiced can backfire, if one doesn't go about it in just the right way (Legault et al. 2011). Reflecting on past instances in which one managed to do the right thing makes one *more* likely, not less likely to be biased (Moskowitz and Li). So what should one do?

Fortunately, there are some things we can do. Obviously anonymising can prevent us from even being aware of the social group that might trigger our implicit biases.[10] But anonymising is not a solution that's always available or appropriate, so it's fortunate that psychologists are discovering a lot of surprising interventions that seem to reduce the influence of implicit biases. We can spend time thinking about counter-stereotypical exemplars (members of stereotyped groups who don't fit the group stereotypes).[11] We can carefully form implementation intentions—not 'I will not be influenced by race' but 'when I see a black face I will think 'safe'' (Stewart and Payne 2008). We can spend a few hours engaging in Kawakami's negation training, in which we practice strongly negating stereotypes (Kawakami et al. 2000). But this might not work, unless we use Johnson's (2009) variant in which we think 'NO, THAT'S WRONG!' while pressing a space bar whenever presented with a stereotypical pairing. We can reflect on past instances in which we *failed* in efforts to be unbiased, thereby activating our motivation to control prejudice (Moskowitz and Li). And these are just a few examples.

Interestingly, some very effective interventions—like Kawakami's negation training—are widely viewed as far too demanding for widespread adoption. Alex Madva (manuscript), however, has argued extremely compellingly that these have been dismissed far too quickly. And he has a point—what's a few hours of slightly tedious exercises if it can actually make me less prejudiced? The arguments I have presented here suggest that we may well also have very strong *epistemic* reasons as well for adopting these techniques. If we don't try to overcome the pernicious influences of these biases, we are not being responsible enquirers.[12]

Importantly, though, we are unlikely to completely eliminate the threat of error. Implicit bias could be affecting one's reasoning at almost any point—it is very hard to judge when social group membership is having a pernicious influence. So it is much trickier to correct for than other factors that are known to make one unreliable (e.g. 'don't make important decisions when drunk'). If we knew that we were about to enter a situation in which implicit biases might impair our thinking, and we knew exactly which biases would be relevant, we could formulate appropriate implementation intentions, like 'If I see black person, I will think "safe". But we don't in general know which stigmatized social groups we will encounter at which points, or what stereotype will be relevant. (Thinking 'safe' when we see a black person will not help us to more accurately assess the quality of their written work.) Moreover, we don't know what sorts of cognitive task might be relevant. So far, I have focused mostly on assessments of quality of argument, or of believability. But implicit biases surely affect other epistemically relevant matters as well: they might lead me to ask the wrong questions, or to neglect the right ones. Implementation intentions are a powerful device for controlling the expression of biases, but by their nature they target very specific behaviours. They cannot provide the general sort of reshaping of the cognitive faculties that would be needed to fully combat the influence of implicit biases. At the end of the paper, I'll discuss what this limitation to our individual corrective measures means for us.

4 Our Rational Capacities

Miranda Fricker is one of the few epistemologists who has thought long and hard about the negative epistemic effects of stereotypes. Her focus, however, is on the way that these affect evaluations of testimony from those that the stereotypes target, and she does not discuss the literature on implicit bias. This literature (as we have seen) shows the pernicious epistemic

influence of stereotypes to extend far beyond evaluation of testimony. Still, Fricker's discussion is highly relevant: she argues that those who underrate the testimonies of others due to wrongful stereotyping of their social group are committing an injustice, and that they suffer from an epistemic vice.* This terminology seems wholly appropriate to apply to those in the grip of pernicious implicit biases. It seems worth examining, then, what she says about correcting for prejudices.

Fricker suggests that there are two ways to be a virtuous agent in terms of accepting testimony. The first is to be 'naively' virtuous—to simply have credibility judgments that are not influenced by prejudice. She admits that this will be difficult to manage with respect to the prejudices of the culture/sub-culture one grows up in. The next is to reflectively correct one's judgments—to, for instance, think 'I'm white, and I may fail to give sufficient credibility judgment to black people as a result.' Or, alternatively, to notice that despite consciously believing women to be the equals of men, one tends to always take a man's word over a woman's. Noticing these things, she suggests, allows one to consciously raise the credibility one assigns to members of stigmatized groups. And this possibility, she suggests, is essential to our status as rational enquirers:

> 'The claim that testimonial sensibility is a capacity of reason crucially depends on its capacity to adapt in this way, for otherwise it would be little more than a dead-weight social conditioning that looked more like a threat to the justification of a hearer's responses than a source of that justification.' (84)

Extending this idea in a natural way, we would expect the capacity to consciously, critically, reflectively correct for one's biases quite generally to be crucial to one's epistemic capacities being capacities of reason.

Before we learn about implicit bias and what to do about it, it is genuinely unclear to me whether we have this ability to critically and reflectively correct for our bias. We could perhaps claim that we had the *ability* to do that (once we learned about the evidence, etc.) but this claim would be so weak as not to amount to actually be very reassuring. Now, however, many of us do have the ability to critically and reflectively correct for our biases—at least once we have learned about their existence and studied the literature on what to do about them. Once we do that (and implement these techniques), we can responsibly claim that these capacities are not just dead-weight social conditioning. Importantly, though, this requires more than what Fricker imagined in her discussion: we are unlikely to notice through individualistic reflection the ways that our judgments are affected by social categories; and even when we do notice this we are unlikely to hit upon the right strategies for fighting it. The only way that we can engage in the necessary sort of correction is not individualistically or introspectively, but by informing ourselves about what scientists have discovered about humans like ourselves. The correction is dependent not just on our rational faculties but on the deliverances of science.

In order to inquire responsibly, we must instead recognize that our epistemic capacities are prone to errors that we cannot learn about through first-person reflection; and that we must correct them using counter-intuitive mechanical techniques that draw not upon our rational agency but upon automatic and unconscious responses. We can consciously enlist these unconscious responses, and use them to improve our epistemic responses, but we cannot do this through rational and critical reflection alone.

Moreover, as I noted in the previous section, individual efforts are inevitably limited.

To fully combat the influence of implicit biases, what we really need to do is to re-shape our social world. The stereotypes underlying implicit biases can only fully be broken down by creating more integrated neighborhoods and workplaces; by having women, people of colour and disabled people in positions of power; by having men in nurturing roles; and so on. The only way to be fully freed from the grip of bias-related doubt is to create a social world where the

---

* Saul describes Fricker's position in terms of virtue and vice, terminology that is often used in contemporary epistemology. An epistemic vice is a character trait that leads to bad knowledge practices; gullibility, dogmatism, and closed-mindedness are common examples, whereas conscientiousness and open-mindedness are epistemic virtues.

stereotypes that now warp our judgments no longer hold sway over us. And the way to do this is to end the social regularities that feed and support these stereotypes. Can this be done? Who knows. It is a massive task—one whose importance and magnitude Elizabeth Anderson makes clear (for the case of race) in her *The Imperative of Integration*. But if it is not, we would seem to be stuck with bias-related doubt, and with the consequent lack of trust in our cognitive faculties. And this is in itself quite a fascinating result. Scepticism is generally thought of as a highly individualistic epistemic issue. It's about the would-be knower doubting the guidance of her own mind. But bias-related doubt shows us a social dimension to this. We have seen that the social world gives rise to a powerful form of doubt, and one that can only be fully answered by a sweeping and radical transformation of our social world.[13] ■

Suggestions for Critical Reflection

1. The main conclusion of this paper is that bias-related skepticism is far more worrisome than traditional philosophical skepticism. In what ways is bias-related skepticism said to be worse than traditional skepticism? Do you agree? What are the main components of Saul's argument for this conclusion?
2. Saul stresses that bias-related doubt is "not *just* a matter of what Miranda Fricker has called testimonial injustice." What does she mean by this? In what way does bias-related doubt go above and beyond testimonial injustice?
3. "The only way that we can engage in the necessary sort of correction is not individualistically or introspectively, but by informing ourselves about what scientists have discovered about humans like ourselves." Why does Saul say this, and how important is it? Is she right?
4. To what extent are people morally responsible for their own biases? Is society on the whole responsible? What are the implications of off-loading responsibility for biases from individuals onto society as a whole?
5. Have you observed implicit bias affecting your own judgment? If you have, how might you attempt to eliminate those biases? If you have not, how might you go about determining whether you are affected by implicit biases?

Notes

1 See, for example, Scheman 2002.

2 For a great deal more precision about the many different ways of characterizing implicit bias, and the many sorts of implicit biases there are, see Holroyd and Sweetman (forthcoming).

3 See, for example, Bertrand and Mullainathan 2004; Rooth 2007; Moss-Racusin et al. 2012; Steinpreis et al. 1999.

4 See, for example Correll et. al. 2002, 2007; Greenwald, Oakes, & Hoffman 2003; Payne 2001; Plant & Peruche 2005.

5 For much more on how perception is affected, see Siegel 2013.

6 Here I am assuming that philosophers will be prone to the same sorts of errors as others. They have not actually been studied.

7 I am *not* saying that we are affected only by biases. Of course, a part of what we are doing is applying our skill in evaluating philosophy, and sometimes we will get things right. My claim is just that these judgments will often be distorted, to a variable extent, by biases.

8 I phrase it this way because what affects us as audiences is what social group we *take* the speaker to be a member of, not what social group they are actually a member of.

9 Hookway (1990: 6).

10 This worked beautifully with orchestras, which began holding auditions behind screens, dramatically increasing their percentages of female members. And it is now standard practice in the UK to mark students' work anonymously, which is supported by the Union of Students for just this reason: <http://www.nusconnect.org.uk/campaigns/highereducation/archived/

learning-and-teaching-hub/anonymous-marking/>. For research on anonymous marking see Bradley 1984, 1993.

11 Blair 2002; Kang and Banaji 2006.

12 Madva also responds to criticisms that these techniques are not effective enough, and that they are too individualistic, focusing as they do on individual thinkers rather than societal reform.

13 I had very useful discussions of this paper with several different audiences: The ENFA5 Conference in Braga, Portugal; the Eastern APA audience in Washington DC; and the departmental seminars at Nottingham and Southampton. I have also benefitted enormously from discussions with Louise Antony, Ray Drainville, Miranda Fricker, Teresa Marques and especially Chris Hookway—to whom this paper owes an obvious and enormous debt. (Though the errors are all mine.)

References

Anderson, E. 2010. *The Imperative of Integration*. Princeton: Princeton University Press.

Beebee, H. and Saul, J. 2011. *Women in Philosophy in the UK: A Report*, published by the British Philosophical Association and the Society for Women in Philosophy. (<9-08/Women%20in%20Philosophy%20in%20the%20UK%20(BPA-SWIPUK%20Report).pdf>)

Bertrand, M. and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? *American Economic Review*, 94, 991–1013.

Blair, I. 2002. The Malleability of Automatic Stereotypes and Prejudice. *Personality and Social Psychology Review*, 3, 242–261.

Bradley, C. 1984. Sex bias in the evaluation of students. *British Journal of Social Psychology*, 23: 2, 147–153.

Bradley, C. 1993. Sex bias in student assessment overlooked? *Assessment and Evaluation in Higher Education* 18:1, 3–8.

Correll, J., Park, B., Judd, C., & Wittenbrink, B. 2002. The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83, 1314–1329.

Correll, J., Park, B., Judd, C., Wittenbrink, B., Sadler, M.S., & Keesee, T. 2007. Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, 92, 1006–1023.

Frances, B. 2005. *Scepticism Comes Alive*. Oxford: Oxford University Press.

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

Goldin, C. and Rouse, C. 2000. Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians. *The American Economic Review*, 90:4, 715–741.

Greenwald, A.G., Oakes, M.A. and Hoffman, H. 2003b. Targets of discrimination: Effects of race on responses to weapons holders. *Journal of Experimental Social Psychology*, 39 399–405.

Holroyd, J. and Sweetman, J. Forthcoming.* The Heterogeneity of Implicit Bias. In *Implicit Bias and Philosophy*, ed. by M. Brownstein and J. Saul. Oxford: Oxford University Press.

Hookway, C. 1990. *Scepticism*. London: Routledge.

Hookway, C. 2000. *Truth, Rationality, and Pragmatism: Themes From Peirce*. Oxford: Oxford University Press.

Hookway, C. 2003. How to Be a Virtue Epistemologist. In *Intellectual Virtue: Perspectives from Ethics and Epistemology*, ed. by M. DePaul and L. Zagzebski. Oxford University Press.

Hookway, C. 2010. Some Varieties of Epistemic Injustice: Response to Fricker. *Episteme* 7:2, 151–163.

Johnson, I.R. 2009. *Just say 'No' (and mean it): Meaningful negation as a tool to modify automatic racial prejudice*. Doctoral dissertation, Ohio State University.

Kang, J. and Banaji, M. 2006. Fair Measures: A Behavioral Realist Revision of 'Affirmative Action'. *California Law Review* 94, 1063–1118.

Kawakami, K., Dovidio, J.F., Moll, J., Hermsen, S., and Russin, A. 2000. Just say no (to stereotyping): effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, 78, 871–888.

Legault, L., Gutsell, J., and Inzlicht, M. 2011. Ironic Effects of Antiprejudice Messages: How Motivational Interventions Can Reduce (But also Increase) Prejudice. *Psychological Science* 22(12), 1472–1477.

Madva, A. 2013. The Biases Against Debiasing. Paper presented at *Implicit Bias, Philosophy and Psychology Conference*, Sheffield, April 2013.

Moskowitz, G. and Li, P. 2011. Egalitarian Goals Trigger Stereotype Inhibition: A Proactive Form of Stereotype

* Now published: 2016.

Control, *Journal of Experimental Social Psychology* 47, 103–16.
Moss-Racusin, C., Dovidio, J., Brescoll, V., Graham, M., Handelsman, J. 2012. Science Faculty's Subtle Gender Biases Favor Male Students. *PNAS* 109(41), 16395–16396.
Payne, B.K. 2001. Prejudice and perception: The role of automatic and controlling processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, 81, 181–192.
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.
Plant, E.A. and Peruche, B.M. 2005. The consequences of race for police officers' responses to criminal suspects. *Psychological Science*, 16, 180–183. *Price-Waterhouse v. Hopkins*, 109 S. Ct. 1775. (1989).
Rooth, D. 2007. Implicit discrimination in hiring: Real world evidence (IZA Discussion Paper No. 2764). Bonn, Germany: Forschungsinstitut zur Zukunft der Arbeit (Institute for the Study of Labor).
Saul, J. Forthcoming. Implicit Bias, Stereotype Threat and Women in Philosophy. In *Women in Philosophy: What Needs to Change?*, ed. by F. Jenkins and K. Hutchison. Oxford: Oxford University Press. (Formerly titled 'Unconscious Influences and Women in Philosophy'.)*
Scheman, N. 2002. Though This Be Method, Yet there Is Madness in It: Paranoia and Liberal Epistemology. In *A Mind of One's Own: Feminist Essays on Reason and Objectivity*, ed. by L. Antony and C. Witt. Cambridge, MA: Westview.
Steinpreis, R., Anders, K., and Ritzke, D. 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, 509–528.
Siegel, S. 2013. Can Selection Effects on Experience Influence Its Rational Role? *Oxford Studies in Epistemology* Vol. 4: 240–270.
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, 1332–1345.
Unkelbach, C., Forgas, J., and Denson, T. 2008. The Turban Effect: The Influence of Muslim Headgear and Induced Affect on Aggressive Responses in the Shooter Bias Paradigm. *Journal of Experimental Social Psychology* 44:5, 1409–1413.

Taken from: *The Broadview Introduction to Philosophy*, Andrew Bailey, Ed. (Ontario: Broadview, 2019).

* Now published: 2013.