sciendo

(How) Should We Tell Implicit Bias Stories?

Jennifer Saul
University of Sheffield

DOI: 10.2478/disp-2018-0014 BIBLID [0873-626X (2018) 50; pp.217–244]

Abstract
As the phenomenon of implicit bias has become increasingly widely known and accepted, a variety of criticisms have similarly gained in prominence. This paper focuses on one particular set of criticisms, generally made from the political left, of what Sally Haslanger calls "implicit bias stories"—a broad term encompassing a wide range of discourses from media discussions to academic papers to implicit bias training. According to this line of thought, implicit bias stories are counterproductive because they serve to distract from the structural and institutional factors that underlie oppression of social groups. This paper argues on the contrary that implicit bias stories, properly told, can help direct attention and concern to structural and institutional factors, and indeed may be especially helpful in motiving action. The key, however, is to tell these stories properly. When implicit bias stories are told in the wrong way, they are indeed counterproductive. This paper looks in detail at several examples of good and bad implicit bias stories, examining what makes some of them counterproductive and others highly effective in motivating action to combat social injustice.

Keywords
Implicit bias, structural injustice, racism, sexism, individualism.

Over the last decade, implicit bias has become an increasingly mainstream topic, not just in philosophy but in the world outside academia. Perhaps unsurprisingly, criticisms of implicit bias discourses have also become increasingly common. My focus here will be on one specific set of criticisms from the political left. These critics argue that implicit bias discourses are counterproductive for the goal of making progress toward social justice, in particular because they distract from the structural and institutional elements of racism that

play such a powerful role in our world. Although I disagree with the sweeping condemnations some have made, I have become increasingly convinced that implicit bias discourses can also sometimes be immensely problematic. This paper represents an attempt to grapple honestly with the criticisms that have been made of implicit bias discourses, and to think through what makes such criticisms appropriate or inappropriate, with respect to various different kinds of implicit bias stories. (I will not be addressing the empirical critiques that have been made of implicit findings in psychology. These are important to address, but they are a large topic for another paper. I take them to be less powerful than the popular narrative on the subject would suggest, for the reasons given in Brownstein et. al. (under review).

A further and related criticism, very salient now, is that the global rise of the far right shows that our energies now need to be focused on the resurgence of absolutely explicit racial hatred and misogyny. I think this criticism has a great deal going for it, but that it does not mean that there is no place for discussions of implicit bias. This paper was initially written at a more hopeful time (early 2016), and is something of a relic of those times. However, there are also local contexts in which explicit hatred has not been normalized and in which it is right to focus on the implicit in the ways discussed here. And it is also worth noting that even where explicit bias is normalized, implicit bias may still have an important role to play. Moreover, some of the methods proposed for dealing with implicit bias (procedural interventions, for example) can also aid in dealing with explicit bias—and may be easier to persuade people to adopt.

I will begin by sketching some key criticisms made by Sally Haslanger and others, of what Haslanger calls 'implicit bias stories'. 'Stories' as I will use the term, is meant in a very broad sense—and can encompass not just narratives, but also training sessions and philosophical writings. Then I will examine good and bad implicit bias stories in three areas: implicit bias training, discussions of police shootings of people of colour, and philosophical writing. From this, I will extract key features that help to make an implicit bias story a good implicit bias story.

First, a word on methodology. Haslanger and I both share the goals of Critical Social Theory, as Haslanger understands it.

> Critical Social Theory begins with a commitment to a political movement and its questions; its concepts and theories are adequate only if they contribute to that movement. (Haslanger 2012: 22)

Haslanger's concerns about (some) implicit bias stories, then, are rooted in the worry that they will not contribute to feminist and anti-racist projects, or that—worse yet—they might hamper such projects. As we'll see, I agree with her that some of them are flawed in these ways. However, I argue that—properly told—implicit bias stories can very much help us in our feminist and anti-racist projects.

A key to my defense of (some) implicit bias stories begins from what Haslanger says about one way that a critical theory can fail:

> The theory does not receive reflective endorsement even after opportunities for reflection have been offered…This suggests that the theory is missing a bridge that allows a shift from seeing the world in one way, to seeing it in another. (Haslanger 2012: 28)

I will suggest below that implicit bias stories, told in the right way, can (in at least some cases) provide just the bridge needed to pull us out of the purely individual perspective that is so profoundly tempting.

1 Implicit bias explanations: Haslanger's concerns and some initial thoughts

Importantly, Haslanger does not condemn all discussions of implicit bias. Instead, she seeks to show some key ways that implicit bias stories can be inadequate. Implicit bias stories can serve to prop up and perpetuate the "ideology of individualism"—an ideology that serves to obscure the structural elements of injustice that urgently need to be both understood and fought. As a part of making this point, and showing its importance, she demonstrates that there are racial and gender injustices which *must* be understood on a structural level—which simply cannot be understood if we confine ourselves to thinking about individual biases, whether explicit or implicit. She also offers some guidance regarding just what it is that makes some implicit bias stories inadequate, and what is required for a more adequate story.

1.1 Standard and nouveau stories

Haslanger contrasts the sort of structural story she thinks we need to tell more of with what Tilly calls 'Standard Stories', and with what she calls 'Nouveau Stories'. Here is Tilly's summary of the elements of a Standard Story:

> (1) limited number of interacting characters, (2) limited time and space, (3) independent, conscious, self-motivated actions, (4) with the exception of externally generated accidents, all actions resulting from previous actions by the characters. (Haslanger 2015: 3)

Haslanger also discusses what she calls 'Nouveau Stories'. While a standard story of inequality between groups focuses exclusively on explicit prejudice, a Nouveau story instead adds implicit bias to the mix. Otherwise, it remains the same. They key point is that it explains systematic injustice *solely* in terms of individual biases, whether explicit or implicit, and does so without any influence (except "acts of God") from forces beyond the individuals who are the story's focus.

Haslanger links Standard Stories to the Ideology of Individualism, and I read her as also linking Nouveau Stories to this. Certainly, I have often encountered (in conversation) the idea that implicit bias stories are problematically individualistic. Haslanger takes a further problem with standard stories (and presumably Nouveau ones as well) to be how deeply the grip us.

> The second problem with Standard Stories is, in a way, their power to capture our imaginations. Standard stories may be irresistible for humans. They are one way we create and reproduce social meanings. (Haslanger 2015: 9)

The thought here is that Standard Stories are incredibly appealing to us, sucking us in and helping to prop up the Ideology of Individualism that prevents us from seeing the structural forces that produce and perpetuate social inequality, and that must be fought if we are to have any hope of arriving at a more just world. This concern is shared by Banks and Ford (2009: 1), who write that "the goal of racial justice efforts should be the alleviation of substantive inequalities, not the eradication of unconscious bias. Yet, the rhetoric of implicit bias is so compelling that people are likely to accept it as the goal of racial reform and, consequently to push the theory in directions that

siphon energy away from problems of substantive inequality…".

I think the most useful way to tackle this topic is via detailed examples, but before we do that I want to flag up a small concern about characterizing implicit bias stories as Nouveau stories. A Nouveau story, according to Haslanger, is one that invokes implicit bias in order to explain systemic inequality. It is otherwise just like a Standard Story. What makes Standard Stories—and so, presumably, Nouveau stories—so deeply inadequate is the total lack of attention to structural matters and exclusive focus on individuals. But this characterization is problematic: many, if not most, implicit bias stories *do* discuss societal racism. They talk about the racist (and sexist, and so on) structure of society as an absolutely crucial background factor that helps to produce individuals who are implicitly biased despite what are often genuine efforts to be unbiased. If this is right then many implicit bias stories will not qualify as Nouveau stories.

Importantly, however, this terminology is not so important. Our focus here is on strengths and weaknesses of various stories that do or don't include discussion of implicit biases. My goal is not just to defend the idea that implicit bias stories can sometimes be helpful, but to also understand when and why they are counterproductive. This will contribute to working out when and how we should tell implicit bias stories.

1.2 Haslanger: structural injustice not explainable in terms of bias

One strong reason to move beyond Standard and Nouveau stories is the existence of social injustices that cannot be understood solely in terms of the biases of individuals. Haslanger gives several compelling examples of this sort. There is Jason, who becomes unemployed when the city cuts the funding to the bus route that takes him to work. And there are a lot of other Jasons, in poor neighbourhoods, who suffer from the same lack of resources as cities decide where to distribute funds. Nobody in the present day need ever make any individually biased decision to perpetuate the inequality that has such a devastating effect on Jason's life. There are also cases like that of Rashaan, Jamal, and their teacher. The teacher in the example is

fairly applying her exceptionless rule of dismissing students who are disruptive when she sends Rashaan out of the room. And yet, thanks to the stigma attached to blackness in their culture, Rashaan and Jamal are fully reasonable in taking this to be an example of racism. Haslanger writes:

> The problem is that in a particular setting, the participants may not themselves hold biased attitudes. Their responses to each other may be based on good evidence; and they may be reasonably interpreting the other in light of the social meanings of the actions that have been performed. The source of the problem, I submit, is cultural rather than individual. (Haslanger 2015: 7)

Haslanger is absolutely right that these problems are cultural. Any attempt to make sense of these stories in a purely individualistic manner will fail.

But Haslanger is also right to note, as she does, that a *purely* structural explanation is also doomed. It is important to take note of the interactions between different levels of explanation.

> For example, stigmatizing meanings generate mistrust that alienate nonwhite teens from school; the lack of education and concern with professional success reinforces the stigmatizing meanings. The optimistic take is that the violent appropriation of the means of production may not be necessary in order to bring about social change (!), for resistant agency and countercultural movements make a difference. Drawing attention to and correcting implicit bias can be part of this effort; but without structural change, cultural contestation, and redistribution of resources, the biases will persist and the most profound injustices will remain entrenched. (Haslanger 2015: 8)

And we can see this interaction in Haslanger's anecdotes. While a purely individualistic story could not fully capture what is taking place, neither can a purely structural one. And it's worth expanding on this a little.

When the decision is made to adopt transportation policies that differentially harm particular groups of people, these may or may not be directly based on biases (whether explicit or implicit) against these people—they could be based on simple ignorance about the importance of public transportation to the lives of people living in a particular area. However, it is very unlikely that biases have played

no role whatsoever in these decisions being made. After all, decisions like this are rarely made without some push-back. Citizens will have complained about the effects on their lives, and someone will have to have decided their concerns were not worth much attention. Economic policies which favour (for example) low corporate taxes over investment in services will have been adopted. And the decision to prioritise in this manner will have been based on biases about who matters, and (again) whose voices to listen to. Further back, there will have been decisions made about who to vote for, what to teach in schools, what to publish in newspapers, and so on. Individual biases, either explicit or implicit, will be a part of these. The homogeneity of the committees making decisions will have been shaped in part by biases. And biases play a crucial role in the credibility deficits and surpluses guaranteeing that some people's voices are heard more easily than others (Dotson 2014; Fricker 2007; Medina 2012; Saul 2017). And these help to hold in place the structural injustices Haslanger describes.

It is even easier to see the individual element in cases like that of Rashaan, Jamal, and their teacher. Note, first, that we cannot see the teacher as fair and *unbiased* without thinking about her as an individual. Next, we are thinking about individual biases when we think of Rashaan and Jamal's past experiences with prejudice. And, of course, we are considering the individual when we consider their beliefs about their teacher's biases.

My point here is to emphasise the importance of individuals and implicit bias to a full understanding of even the structural explanations Haslanger gives. Her cases are indeed cases where implicit bias stories on their own would be inadequate. But they are also cases in which a structural explanation on its own is incomplete. Haslanger, as I've noted, does not claim that implicit biases will play no role in these cases. Nonetheless, I think it is worth bringing out the wide range of sometimes hidden ways in which implicit biases can prop up structural injustices.

2 Implicit bias training: criticisms and my responses

Implicit bias training has become increasingly widespread both in the corporate world and in academia. This is, to my mind, a great

advance over either total neglect of these issues or an exclusive focus on conscious, deliberate, and explicit bias. Yet implicit bias training has also been subject to a great deal of criticism; and much of this, we will see, I take to be quite appropriate. Nonetheless, I will argue that there are important roles for implicit bias training to play, if carried out in the right way.

2.1 Criticisms

We'll start with a popular presentation of empirical criticism. A November 2015 article from Fortune magazine (http://fortune.com/2015/11/10/test-racism-sexism-unconscious-bias/) cited "one study [that] even found that the training can have negative effects, validating individuals' biases and thereby perpetuating them." The study cited here (Duguid and Thomas-Hunt 2015) is an interesting one. It shows that raising people's awareness of the pervasiveness of implicit bias *increases* rather than decreases their tendency to make biased decisions. They take this to be due to a kind of moral liscensing that comes from the realization that bad behavior is widespread. This has been noticed in other areas: learning about the extent of problems with littering might increase the propensity to litter, but learning that the overwhelming majority of people in your neighbourhood never litter is likely to decrease it. The thought here is that we tend to conform to the norms that we believe to be prevalent. So learning that biased behavior is widespread will make us more relaxed about being biased. A key component of implicit bias training, normally, is making participants aware that everyone is subject to implicit bias, in order to make participants aware that they are likely to be biased. So this study might seem to be quite devastating—it seems to show that implicit bias training is not only ineffective, but counterproductive.

In work not yet published, Joy-Gaba (Dissertation 2011) studied the effects of implicit bias training sessions that she herself designed and implemented. Her research found that this training increased *knowledge* about implicit bias, causing participants to be more likely accept, for instance, the likelihood that they were biased; but it did not have any effects on tendencies to make biased decisions. Since the point of implicit bias training is not just to impart information

but to change behavior, this again looks very worrying for implicit bias training.

The main worry that Haslanger raises about implicit bias stories is their individualism. Her thought is that they lead us to neglect the structural factors involved in inequality and oppression. This is a different worry from the empirical ones above. The empirical worries raised above are not ones that stem from a concern with individualism. The studies discussed involve training individuals and testing for the effects of this training on these individuals. They show the training to have been either ineffective or counterproductive. But they do not break out of the individualistic paradigm[1]. Even if they had shown individuals to display reduced bias levels, the worry raised by Haslanger would remain. The concern, I take it, is that if we focus on training individuals we may achieve some degree of progress; but we do nothing to change the structural factors—housing and employment segregation; laws relating to childcare leave; all-white syllabi; and so on—that firmly hold in place an inegalitarian world. Even if the training worked by its own lights, this critique goes, that would not be enough.

2.2 A closer look at the inadequacy

I think all of the criticisms do indeed apply to some implicit bias training.

But they do not apply to all implicit bias training. Thus, it seems to me, they should be seen as problems for *poorly designed* implicit bias training. While it is not impossible that they might apply to better designed implicit bias training, that cannot be said to be established.

Let's begin with the empirical critiques, and take them individually. Duguid and Thomas-Hunt (2015) did indeed show that telling people about the pervasiveness of implicit bias increased the manifestation of implicit bias in behavior. But the studies which got this result involved telling people about this pervasiveness *and doing nothing else*. These don't tell us anything about training which is not of

[1] It is, however, worth noting that this paradigm is not *purely* individualistic. As noted earlier, one cannot make any sense of implicit bias in a *purely* individualistic manner.

this form. And in fact, they also ran studies (reported in the same paper) in which participants were told that most people work very hard to overcome their biases—a condition designed to highlight a prevailing norm against being biased. Participants in this condition displayed lower levels of bias. So the results which had been taken to show the ineffectiveness of implicit bias training really only showed the ineffectiveness of one particular form of implicit bias training, while another study *in the same paper* showed that a different kind of training could be more effective.

A similar criticism applies to the training that Joy-Gaba discusses. The training sessions she tested did not simply inform about pervasiveness, but went much further in imparting a solid understanding of the workings of implicit bias and the problems causes by it. However, they also failed to teach specific techniques for overcoming implicit bias. Once more, this seems to me to count as poor implicit bias training. We should not be shocked that if we fail to teach people how to combat implicit bias they do not overcome it.

I have been amazed at how common it is for implicit bias training to do just what has been shown to fail in these studies. Anecdotally, I hear report after report of training sessions in which people are told about the widespread existence of implicit bias and given no instructions at all in how to overcome it or prevent its operation. We already know (from, for example, the papers just cited) this will make implicit bias worse. A "driver training" course that teaches people about the pervasiveness of bad driving and the various accidents it could cause—without teaching people who to drive well—would never be accepted as driver training. I think we should take a similar attitude to an implicit bias course that teaches about the pervasiveness of implicit bias without teaching how to overcome it.

Many actual training sessions will also send people away telling them that just knowing about implicit bias will make them unbiased; or that they should overcome it by simply striving to be unbiased; or that they should try to be objective. We already know from the literature on implicit bias that all of these instructions will actually worsen the problem (see, for example, Uhlman and Cohen 2006). Training like this is indeed worse than useless, and it's very common.

A well-designed implicit bias training can do much more than simply tell people that implicit bias is pervasive or give them brief

instructions that worsen the problem. Given what we know about the mechanics of fighting implicit bias, a good implicit bias training should also (a) make it clear to people just what damaging effects implicit bias has on the world, so that they are motivated to try to overcome it; and (b) equip people to combat implicit bias by teaching them techniques for doing so. Importantly the techniques taught in implicit bias training may include not just methods that individuals can employ to try to reduce their manifestation of bias, but also structural and procedural interventions. I'll treat the structural and individual separately, for reasons that will become clear.

2.2.1 Implicit bias training can teach individual interventions

There is a significant body of evidence showing the potential for interventions carried out by individuals to reduce level of implicit bias, as measured by IATs and other tests. One particularly notable study is Devine et. al. (2012), which taught participants a collection of bias reduction strategies and measured the effects over a twelve week period. In this study, participants who were taught the interventions showed a significant reduction in implicit race bias over the period, and those in the control group did not. This study appears to show that effective bias-reduction techniques can be taught. If that's right, then this is something that implicit bias training can and should be doing.

There are, in addition, further individual techniques that seem to effectively reduce levels of implicit bias. These techniques are somewhat time-consuming and labour-intensive, so much so that many have dismissed them as impractical. However, Alex Madva (2017) has argued that—given the substantial real-world effects of implicit bias—we should take these remedies far more seriously than most have tended to do. If that's right, then these also can and arguably should be taught.

However, it is important to note that these remedies are limited in important ways. Even at their best, they merely show a reduction in levels of implicit bias, rather than its elimination. Moreover, the remedies Devine et. al. studied are most effective in those who are concerned about discrimination—and not everyone is. Similarly, the labour-intensive remedies Madva discusses will simply not be carried out by those who are not strongly motivated to reduce prejudice.

Finally, a more recent study by Devine and colleagues (Carnes et. al. 2015)—this one of gender bias—did not show reduction in IAT levels (at least for the association that they tested, of gender and leadership).[2] This at least raises questions about the efficacy of the techniques taught by Devine, though it by no means settles the issue. All of these considerations suggest a need for training not to focus just on bias reduction strategies for individuals.

2.2.2 Teaching structural and cultural interventions in implicit bias training

Implicit bias training need not, however, focus solely on bias-reduction strategies for individuals. Those that do, however, would rightly be subject to Haslanger's criticism: we should not focus just on individualistic remedies. Doing so is of only limited effectiveness, and Haslanger is right to suggest that it would also encourage a view of the world on which systemic inequality is wrongly understood in purely individualistic terms, and solutions are sought solely at the level of individual behavior.

What would implicit bias training be like that discusses more structural solutions? Many of us who do research in this area are already offering such training. We discuss interventions such as the following, adapted for appropriateness to institution. (I offer just a few examples.)

(1) Anonymising grading, CV review, writing sample review, and so on—in order to remove opportunities for bias to affect evaluation.

(2) Altering discussion conventions so that they are not so dominated by, for example, older white men. In my trainings I suggest making sure to look out for contributions from people who are not usually heard from, calling attention to good contributions that are being overlooked or misattributed, and using the One Question Per Question rule in running seminars.

[2] This might be viewed as a failure to replicate the previous study, but it might also indicate some significant difference between interventions needed for gender and race bias.

(3) Reviewing procedures for promotion, hiring, etc. to reduce opportunities for bias.

(4) Seeking ways to integrate counterstereotypical exemplars into organizational life: at a minimum, considering whose pictures hang on the walls; in academia, working on more diverse reading lists and conferences—both in terms of demographics and in terms of content; and, crucially, attempting to diversity the demographics of the workforce.

When I do one-off training sessions, I place substantial emphasis on institutional and cultural change. But I also note the importance of following up on my brief suggestions with further thought and discussion. I offer to discuss ways of reforming culture and procedures with interested parties over coming weeks and months. One recent training session yielded four extensive follow-up discussions, still ongoing, about ways to change procedures. It may well be that those who issued the initial invitation simply wanted me to suggest individual debiasing procedures rather than cultural change. But a well-run bias training session allows one to begin from individual-level concerns and move on to cultural and institutional reform. Importantly, many of the most important discussions and changes will occur as part of the follow-up.

Implicit bias training is in fact ideally suited to starting from individual-level concerns and motivating a move to structural and institutional ones. While it would be theoretically possible to discuss implicit biases as merely surprising facts about individual psychologies, it makes much more sense to explain them as resulting from hierarchically structured societies. Once this is done, we're no longer on solely an individual level.

A further concern one might have about implicit bias training is that learning how widespread and automatic biases are could make people feel absolved from responsibility for them. This sort of liscensing effect may in fact play a role in the backfiring of some of the sorts of training discussed earlier in this paper. However, this need not be the way that one presents implicit biases. I always close by noting my view that merely possessing implicit biases is not always blameworthy, as one might be unaware of this fact or unaware of

what can be done to combat implicit biases. But I also always note that after hearing my talk people now know enough that they would be blameworthy if they chose not to do anything, and I emphasise that I have given them a handout filled with techniques to try.

A particular strength of implicit bias training as a cultural intervention is the way that it moves from the individual to the institutional and cultural. As Haslanger rightly notes, individualistic stories exert a peculiarly strong grip on us. People are slow to acknowledge the role of structural factors, and resist the idea that racism or sexism could be matter of something other than individual, conscious, deliberate bias. My implicit bias training exploits this tendency to be drawn to the individualistic, by making a great deal of the fact that every individual, including those at the training session, is pretty much certain to possess pernicious implicit biases. This grips people, worries them. But, the training continues, this is largely the result of being raised in a culture structured by unjust racial, gender, and other hierarchies—not a sign that the audience member is really a terrible person. This idea is a relief. But to get the relief from embracing this idea, one must also embrace the idea that society is structured by unjust racial, gender, and other hierarchies. All of a sudden people become eager to accept a notion of structural inequality that they might have resisted before. And then solutions are presented. They are presented as a menu from which to choose, but the audience is also made aware that they have a real moral obligation to find *some* way to take action against implicit bias. The solutions are, as noted, both individual and institutional. While the institutional solutions are more powerful and far-reaching, not everyone at the talk will be in a position to implement them. This is why it's important to also offer other options, so that each individual has some action to take, even if that action is simply one of talking to one's colleagues about institutional solutions to be tried. In short, implicit bias training is appealing because it is based in interesting and important facts about individual psychology, and people are fascinated (and disturbed) by such stories. But to understand implicit bias one needs to situate it in an understanding of structural injustice. And many key methods for combatting it are not individualistic, but instead institutional. Implicit bias training, then, forces one to make a move from the individual to the structural, and this is to my mind

one of its great strengths. It can provide the bridge that helps people to move—perhaps without even seeing that this is happening—away from the individualistic view of the world with which they start. And this, as we've noted following Haslanger, is crucial to the success of a critical theory.[3]

A further point worth noting is this: *sometimes* people actually do move quickly to a structural story, and not in a productive way.[4] For example, when confronted with figures about low numbers of black people in Philosophy, people are extremely quick to respond by pointing to residential and educational segregation; society-wide links between race and poverty; and so on. Similarly, a popular response to concerns about low numbers of women in philosophy is to point to "cultural" factors that might make philosophy less appealing to women. What's distinctive about the a*ppealing* structural stories, usually, is that they let their teller off the hook. They offer a response that locates the problem *somewhere else,* somewhere that we can't really be expected to do anything about. As a result, they all too often do nothing to motivate the sort of action that Haslanger and I agree we need to motivate. It seems to me—admittedly this is speculative—that the ideal sort of story for motivating actual action is one that combines the individual and the structural. It needs to make the individual feel a responsibility to *do something,* both individually and structurally. And I think implicit bias stories, told properly, can help us to offer such motivation.[5]

[3] I think it is also important for audiences to be reminded—even if it is not the main subject of the talk—that implicit biases and structural injustices are far from the only things to worry about. I always make a point of reminding audiences of the importance of tackling more overt phenomena like harassment (racial, sexual, and other) and explicit bias, both of which are sadly still very widespread. Some of the measures put in place for implicit bias will help with these—e.g. anonymising may prevent people from acting on explicit biases—but others will require different remedies. I also encourage audiences to follow up with me on these topics.

[4] I thank Jules Holroyd for impressing this point on me.

[5] Banks and Ford (2009) rightly caution that strategically using implicit bias to motivate attention to racial injustice carries risks. Their concern is that if we motivate action by *deceiving* people into thinking our sole goal is eliminating implicit bias, we may find ourselves in a future world shaped solely by this concern

Importantly, different methods will be needed to measure the effectiveness of implicit bias training that emphasizes structural solutions and longer-term consultation on procedural reforms. Measuring individual bias levels via tests such as IATs is not a good way to determine the effectiveness of a session that aims to convince organisations to improve their procedures. Nor would it be a good test of the effectiveness of the procedures that have been put in place. But it will be important to research the effectiveness of this sort of implicit bias training. It seems promising, but to show that it is will require further study.

3 Good and bad implicit bias discussions of police shootings

Thanks to the vital work of the Black Lives Matter movement, police shootings of unarmed people of colour are finally getting attention. Often media coverage and political commentary discusses implicit bias as a cause of these shootings. In my view, this is more often than not done in a way that is counterproductive. Before I explain this, however, it is worth a little bit of background as to why it is appealing (and, sometimes, appropriate) to invoke implicit bias in discussions of police shootings.

The most obviously relevant studies are those of the Shooter Bias. In the Shooter Bias task, participants play a video game in which their job is to shoot quickly if an image of a person with a gun appears and to not shoot if the person is not armed. They are then presented with images of people holding ambiguous objects, that could be either (for example) a phone or a gun. Initial studies, as is common, were with university students. They showed a marked tendency to mistakenly shoot unarmed black but not white subjects. Later studies with police subjects have been more inconclusive, with some

---

and hostile to others. This is why it is, I think, vital, to always situate implicit bias concerns firmly in a story of the structural injustices with which they are intertwined. It is also crucial to call attention to the need for structural reforms to combat implicit bias, demonstrating the extent to which these cannot, practically speaking, be separated. And it is important to remind people—although these days such reminders are less needed than they were when I wrote this paper—that explicit bias is still very widespread.

showing similar patterns, but some quite different.[6] The reasons for these mixed findings are still a matter for debate.

It is easy to see why it is tempting to invoke shooter bias in cases where the shooting involves an officer mistaking a non-gun for a gun, in the hand of a black person. But there are problems even with these, the best cases for implicit bias explanations of police shootings. First, the mixed results with police subjects pose potential problems for these explanations, and could be seen as indicating a need to look elsewhere to understand these shootings (perhaps to explicit bias, or to procedural issues). Even setting this aside, however, there are many ways in which an implicit bias explanation can go wrong. In itemizing a few of these ways, I will be assuming that we are discussing cases in which implicit bias *has* played some role in causing an officer to shoot an unarmed black person.

- Shooter bias is by now well-known, and it has been shown that training can reduce it (Plant and Peruche 2005). If officers are susceptible to shooter bias, lack of proper training is likely to play a role. This needs to be discussed.

- There is, to put it mildly, a significant pattern of police shooting unarmed black people and not being held accountable for it. This lack of accountability makes further shootings more likely. This needs to be discussed.

- Even if implicit bias played a role in, for example, the first shot being fired, all too often these cases involve further shots once the victim is already obviously not a threat, and failure to provide medical attention. These further actions and failures to act are not plausibly justified as due to implicit bias. (It should be clear that one needn't shoot someone who is already incapacitated, *even if* the object they are holding seems to be a gun, and also clear that people in need of medical attention should get it, *even if* they were once holding guns.)

A further problem is that many of these cases are ones in which implicit bias does not seem likely to offer us much of an explanation

[6] For an overview of shooter bias studies, see Cox and Devine 2016.

at all. In the case of Terence Crutcher's shooting, for example, a shooter bias explanation has no appeal: his obviously empty hands were in the air.[7] In the case of Michael Brown, the facts are greatly contested. But even on the version of the shooting told by Derren Wilson, the shooter, there was no thought that Brown was carrying a gun.[8] Wilson's story has it that Brown ran toward him and that he found this very threatening. In neither of these cases was there a mistake about whether the victim was armed. Instead, in both cases the officer thought it was acceptable to shoot an unarmed person. *This*—and the causes for this belief—seems like the primary things that need to be addressed. (And this is so even if implicit bias played some role in causing them to view black men as threatening.)

This belief that it was acceptable to shoot in these circumstances is likely to result at least partially from institutional causes: an absence of proper training, and—crucially—an absence of accountability. If officers were thoroughly trained not just specifically with regard to implicit bias, but also not to shoot at unarmed subjects, and if they were held accountable when they did so, such incidents would be far less likely to occur. A further institutional cause is a tendency to heavily police black areas more than white ones, which will lead to more incidents of this sort occurring with black victims than white—and we know that in Ferguson the heavy policing of black areas was used as a money-making strategy for the government. Another cause is the presence of racist cultures—we know, for example, that Derren Wilson's supervisor sent explicitly racist emails.[9] Although this is partly an institutional matter, it is also very clearly partly a matter of explicitly racist (Standard Story-type) individuals in positions of power. And these are just a few of the additional causes likely to be involved.[10]

---

[7] See, for example, http://heavy.com/news/2016/09/terence-terrence-terance-crutcher-officer-betty-shelby-tulsa-oklahoma-black-man-shot-unarmed-video-family-photos-car/.

[8] See, for example, http://www.independent.co.uk/news/world/americas/michael-brown-shooting-what-happened-in-ferguson-10450257.html.

[9] http://countercurrentnews.com/2015/03/fergson-cop-who-just-got-fired-for-racist-emails-was-darren-wilsons-supervisor/#.

[10] For more on the wide range of problematic issues in Ferguson, see United

So, the full story in any of these cases will involve institutional failures of accountability and possibly training, sometimes combined with implicit or explicit bias. To focus on just implicit bias is to miss what are—from the standpoint of ending these police shootings—the more important institutional factors. Why do I say that the institutional factors are more important? Because they are the ones that will make a difference *no matter what* the causes are in individual cases. Proper accountability and training will make a difference to the manifestation of implicit bias, and will also help to weed out or block the operation of explicit bias. And without these, talking about implicit bias will accomplish nothing. Moreover, in many cases implicit bias may well not be a factor at all.

Talking just about implicit bias also carries the very real risk of exculpating in cases where this is wholly inappropriate. I have been amongst those suggesting that in some cases—where someone is (non-culpably) uninformed about implicit bias and what to do about it, and where implicit attitudes are contrary to genuine explicit ones—it is a mistake to blame people for their implicit biases. But this is assuredly not the case today with respect to police shootings. *Even where* implicit bias plays a significant role in causing a police officer to shoot an unarmed black person, it would be inappropriate to consider the incident a blameless one. The phenomenon of implicit bias, its relevance to police work, and how to combat it, are now extremely well known. Any police department that does not take appropriate measures to fight implicit bias—where these include not just training but also accountability—is very much worthy of blame; and the same is almost certainly true of individual officers, who have a responsibility to educate themselves and guard against the possibility of shooting those they are meant to protect.

Despite all of this, it is *not* always a mistake to discuss implicit bias in the context of police shootings: there are some cases in which it does play a role, when an innocent object is genuinely mistaken for a gun. But *even in these cases* it is an enormous mistake to stop with this. Pointing out the role of implicit bias, where appropriate, should be a way to motivate proper attention to both training and accountability. It is these institutional reforms that are needed in order to

---

States Department of Justice Civil Rights Division 2015.

do something about implicit bias in policing. And, given, knowledge of this, implicit bias must never be used—in this day and age—to excuse police shootings. Each one of these must be viewed as an institutional and individual failure, and attention must be devoted to the institutional actions that can prevent these in the future. Once more, the idea is to use the implicit bias story as a bridge which can motivate attention to the institutional work that is needed.

4 Philosophical implicit bias discussions

4.1 A flawed discussion

Tamar Gendler is the author of one of the earliest philosophical papers on implicit bias, her 2011 "On the Epistemic Costs of Implicit Bias". It's an important and deeply interesting paper, and has been instrumental in helping to bring mainstream philosophical attention to issues related to race. And this furthers the cause of anti-racism, at least within philosophy. Nonetheless, we can see the paper as embodying some of the features that, considered with respect to the goals of critical theory, keep an implicit bias story from being as helpful as it could be.

Gendler's paper explores a fascinating apparent epistemic/moral puzzle. Here's how she summarises the import of the paper:

> if you live in a society structured by racial categories that you disavow, either you must pay the epistemic cost of failing to encode certain sorts of base-rate or background information about cultural categories, or you must expend epistemic energy regulating the inevitable associations to which that information—encoded in ways to guarantee availability—gives rise. (Gendler 2011: 37)

A key case for her is that of a renowned black author who was mistaken for a coat check attendant at a nearly all-white club. The implicit bias that led the woman seeking her coat to make this mistake was based, in Gendler's version of the story, on the wholly accurate generalization that a black person at this club was far more likely to be an attendant than a member, even if they were not wearing the standard uniform. Gendler argues that the rational thing to do is to

encode and act on this base rate information, but that moral considerations pull one in the opposite direction—insisting that one should refuse to take this information into account, in order to avoid acting in a biased manner.

Much of Gendler's paper is not explicitly focused on implicit bias, although that's the official topic of the paper overall. The one section which is most focused on implicit bias discusses an experiment which seems to show that a mismatch between explicit anti-racism and implicit anti-black bias can make it cognitively depleting for many white people to interact with black people. In this experiment, white subjects took an Implicit Association Test, then interacted with either a white or black person, and finally carried out a Stroop task, which is cognitively taxing. White people with implicit anti-black associations had the most trouble with the Stroop task, and also showed high levels of activation in areas of the pre-frontal cortex associated with self-regulation. Gendler writes:

> Together, the neuroimaging and behavioral evidence suggest that participants whose occurrent aliefs[11]—as manifest through their IAT results—were out of line with their conscious goal of acting in a non-discriminatory fashion expended significant cognitive effort to suppress the response-tendencies activated through these associations. (Gendler 2011: 54)

Gender also discusses experiments that seem to show that people refuse to take base rates into account if doing so might seem racist, something known as the Forbidden Base Rate phenomenon. Tetlock and colleagues had volunteers attempt to set housing insurance premiums, taking into account information about actuarial risk, that either did or did not correlate with neighbourhoods' racial composition. They found that subjects were generally unwilling to take this information into account if there was a correlation with neighbourhood racial composition. Gendler writes:

> The phenomenon of Forbidden Base Rates highlights some of the ways in which it is costly to adopt a particular sort of anti-racism in a racially stratified society. It is costly in a narrow economic sense because it causes participants to discount information that, if taken into consideration, would increase their narrowly construed financial well-being;

[11] 'Alief' is Gendler's term for the belief-like structures postulated as at the root of implicit biases.

> and it is costly in an epistemic sense because it causes participants to discount information that might be relevant to their full consideration of both background and foreground conditions. (Gendler 2011: 55)

The overall moral of Gendler's paper is that we find ourselves in a tragic dilemma:

> [The existence of race as a social category[12]] makes encoding information about racial inequity itself problematic—you are faced with a choice between explicit irrationality through base-rate neglect or implicit irrationality through encoding associations that you reflectively reject. (Gendler 2011: 57)

And that's how it ends—on a very depressing, pessimistic note. We are left thinking that there is no way out of this for the foreseeable future, as Gendler takes all this to stem from the very existence of race as a social category. The only way out, it would seem, is to eliminate the category of race. And it's not at all clear how we are meant to do that. (Gendler doesn't propose this, but her view in the paper seems to be that we will be in this dilemma as long as there are races.)

To begin to see the problem with this implicit bias story, it helps to reflect on where it leaves us. It leaves us with a sense of a problem, and no sense at all of a solution. The problem is that we cannot manage to be fully rational and anti-racist.[13] The only way out that is even hinted at is to give up on race altogether, a highly controversial and incredibly difficult to achieve goal. Worse yet, the paper can be seen as (inadvertently) undermining anti-racism movements, by suggesting that opposition to racism leads one into irrationality. After all, on her view, the person committed to anti-racism will—one way or another—fall into irrationality. Although it is clearly not Gendler's intent, this fits exceptionally well with the right-wing narratives of politically correct thought-police attempting to prevent people from facing up to difficult truths; and of the over-emotional left, which

---

[12] Gendler's words, but from slightly earlier.

[13] Interestingly, those who are not anti-racist are immune from the problems Gendler describes. They'll happily take base-rates into account, and they don't reject the associations that they are encoding. Gender does *not* suggest that this is the right way to go. But I think it is significant that this is the only way to avoid the dilemma as she poses it. Egan (2011) notes this as a third option.

really needs to be corrected by the sound common sense of the right. Anything that props up these narratives runs the risk of working against the cause of social justice.

4.2 A better discussion

Now let's contrast this with an alternative take on the same story.[14] Alex Madva (2016) revisits the dilemma, and shows very effectively that there is a way to be both rational and anti-racist. Madva argues that social knowledge— about all of those racial base rates, for example—is not at odds with acting on anti-racism. It is not, Madva argues, social *knowledge* that poses a problem, but rather the use that one may make of this knowledge.

In those cases where implicit biases are the issue (as noted, these are the official but not actual focus of Gendler's paper), the problem is one of chronic accessibility, which leads people to (often automatically) draw on associations when they should not. To combat this, Madva notes that there are many effective suppression techniques that one can use. Madva goes through these in some detail, and argues that individuals should do these things, even where they involve considerable personal effort.

But Madva also looks at other key cases Gendler discusses, like the Forbidden Base Rates. And here he makes a very important point. This is that both the psychologists Gendler draws on, and Gendler herself, assume that participants have the sole goal of maximizing *economic* gain. But, as he notes, the instructions do not tell people what criteria they should use in setting insurance rates, what goals they should have, or what they should maximize. Once racial disparities are mentioned, anti-racists' goals of reducing racial inequalities or compensating for racial injustice may become important to their decision-making. And if these are their goals, they are wholly rational to refuse to base the insurance rates solely on actuarial risks

[14] Madva's discussion is an especially useful one for me to highlight because it is so very focused on actions that can be taken. But another excellent one comes from Katherine Puddifoot, who dissolves the dilemma by arguing that the best ethical choice *is* also the best epistemic choice, because "automatic stereotyping can be poor from an epistemic perspective even if the stereotype that is activated reflects reality" (Puddifoot 2017: 73).

that correlate with race, and to focus instead on reducing race-based inequalities. These structural inequalities are exactly what a rational person concerned with racial inequalities should be thinking about, Madva argues.

Madva's approach is far better than Gendler's from a Critical Theory perspective. Both of his responses look very promising for furthering the cause of anti-racism. With respect to implicit bias, Madva calls upon individuals to take urgent, individual action and gives instructions for how to do this. With respect to Forbidden Base Rates, Madva argues for the legitimacy of allowing political goals to steer one's judgments. In both cases, he firmly rejects the thought that anti-racists are doomed to irrationality. He does not prop up the right-wing narrative about the irrationality of the left, and he offers a road-map of actions to be taken.

Interestingly, however, Madva's approach is still a largely individualistic one. He is focused on showing that individuals don't face a tragic dilemma, justifying the rationality of individual decisions to avoid race-linked insurance premiums, and offering actions that individuals can and should take to combat implicit bias. Despite this, I think his approach does much to serve the aims of anti-racism. Moreover, it helps to combat pernicious ideology—it rejects the linking of leftists with irrationality, and the framing of racism as a problem too overwhelmingly big to act upon. I think Madva offers us a great example of a way in which a broadly individualistic story can still be very useful to critical theory. (It is worth noting, though, that he argues along the way for the appropriateness of *individuals* to be concerned with *structural* inequalities in the insurance rates case.)

But it is also worth thinking about what could be gained by adding a less individualistic approach. Interestingly, such an approach would fit well within the problem as originally set up by Gendler. We could frame the dilemma in just the way that she does, even accepting her understanding of the Forbidden Base Rates case. But rather than simply note the tragedy of it all, one could instead look for solutions. Rather than make a cursory reference to this all following from the existence of race as a category, one who accepted the dilemma *but also* had a commitment to critical social theory would really tackle the issue of what is needed to dissolve the dilemma. Do we really need to eliminate race? If so, why? And how? What would

it mean to do this, and how can we work toward that point? Might there be other collective options to consider? The focus would be on collective action to move beyond the dilemma, rather than just on the dilemma. Importantly, this approach takes anti-racist action to be the key to resolving the dilemma—thus giving it a role in bringing about a world in which one can be simultaneously both rational and ethical. This is in sharp contrast to Gendler's own framing, which makes it all too tempting to see anti-racism contributing to irrationality. This move to a focus on collective action offers us another way, then, to use Gendler's dilemma as an impetus for change.

4.3 What makes the difference

Reflection on different approaches to Gendler's dilemma shows us that one way of getting a better implicit bias story is to move away from the individual level to the social or collective level. But this is not the only way. It seems to me that Madva's approach, which is much more individualistic, also offers us a much better story. A good implicit bias story will be one that furthers the cause of social justice. The traits that are most important to this are the following:

(1) The story situates implicit bias as a result of and contributor to broader structural injustice, and does not underrate the importance of combatting structural injustice.

(2) The story is one on which seeking progress toward social justice is possible.

(3) The story is one on which seeking progress toward social justice is desirable.

(4) The story motivates action (collective or individual) toward social justice.

(5) The story offers a road-map for such action.

Madva's story, despite its individualism, does all this. The collective action story I gesture at above might or might not do all this, depending on how it is fleshed out. It seems to me that if it does not end up

offering a road-map, it will potentially be a less effective story from a Critical Theory perspective, even if it is less individualistic. (Of course, it may be that simply moving people away from an individualistic perspective is so important to the cause that this story ends up being more effective than Madva's individualistic one, despite its road-map. But this is an empirical matter.)

My suggestion is that the elements noted above will, in general, be the ones that make an implicit bias story useful in furthering the cause of justice, and that their absence will make an implicit bias story either not useful, or even counterproductive. Critics are right to worry about some of the implicit bias stories that are being told. But they are wrong to move from that to a broader condemnation of implicit bias stories. Nonetheless, it is vitally important that we take seriously these critiques and make sure that the implicit bias stories we tell are the good ones, rather than the bad ones.

A methodological concern

In this paper, I have examined individual implicit bias stories and their potential to motivate social change. And I remain convinced that this is worthwhile to investigate. But this is not the whole story. Consider, for example, the implicit bias story told by Gendler, which I criticized as likely to motivate passivity or even rejection of social justice movements. There is more that needs to be considered. First, as I alluded to, this paper has also had the effect of (along with others) encouraging mainstream philosophers to begin engaging with philosophy of race, which is leading to increasingly mainstream engagement with increasingly radical work that advocates much more strongly on behalf of social justice movements. It has also inspired an array of responses (e.g. Egan, Madva, Puddifoot,) that seek to respond to and dissolve the dilemma posed. Much of this work is, again, much more clearly motivating of action on behalf of social justice. This raises concerns about the methodology of examining individual implicit bias stories. If a key desideratum for an implicit bias story is that it motivate action, it may be a mistake to look at these stories in isolation. For we cannot really know whether a story will motivate action without a thorough examination of its context and actual consequences. But now another problem arises, once we

take broader context and consequences into account. It is entirely possible for, to take just one example, a very racist story to end up motivating a social justice movement in response. Even though it might have desirable consequences eventually we would not want to praise the story itself. The trick, then, will be to figure out where to draw the line when considering implicit bias stories (or others) in their contexts. This paper can be considered an effort to explore what follows when we consider these stories in a relatively isolated manner. Exactly how much contextualizing we should go in for will be a matter for future research.[15]

Jennifer Saul
University of Sheffield
j.saul@sheffield.ac.uk

References

Banks, R. Richard; and Ford, Richard Thompson. 2009. (How) does unconscious bias matter? Law, politics, and racial inequality. *Emory Law Journal* 58: 1053.

Brownstein, Michael; Madva, Alex; and Gawronski, Bertram. (Under review.) Understanding implicit bias: putting the criticism into perspective.

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* 90 (2): 221–30.

Cox, William; and Devine, Patricia. 2016. Experimental research on shooter bias: ready (or relevant) for the courtroom? *Journal of Applied Research in Memory and Cognition* 5: 236–8.

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.

Dotson, Kristie. 2014. Conceptualizing epistemic oppression. *Social Epistemology* 28(2): 115–38.

Egan, Andy. 2011. Comments on Gendler's 'The epistemic costs of implicit bias'. *Philosophical Studies* 156: 65–79.

Gendler, Tamar. 2011. On the epistemic costs of implicit bias. *Philosophical*

[15] I have had very useful discussions of this paper with Shannon Dea, Ray Drainville, Jack Glaser, Sally Haslanger, Jules Holroyd, Alex Madva, Teresa Marques, participants at NOMOS 2016 (Barcelona), Bias in Context: Psychological and Structural Explanations (Sheffield), the 2016 SWIP Germany Conference (Basel); and the 2018 Implicit Bias and Policing Conference (Sheffield).

Studies 156: 33–63.
Haslanger, Sally. 2012. Introduction. In *Resisting Reality.* Oxford: Oxford University Press, 3–34.
Haslanger, Sally. 2015. Social structure, narrative and explanation. *Canadian Journal of Philosophy* 45(1): 1–15.
Joy-Gaba, Jennifer. 2011. *From Learning to Doing.* PhD Dissertation University of Virginia.
Madva, Alex. 2016. Virtue, social knowledge, and implicit bias. In *Implicit Bias and Philosophy*, ed. by Michael Brownstein and Jennifer Saul. Oxford: Oxford University Press.
Madva, Alex. 2017. Biased against de-biasing. *Ergo* 4(6).
Medina, José. 2012. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations.* New York: Oxford University Press.
Plant, E. A.; and Peruche, B. M. 2005. The consequences of race for police officers' responses to criminal suspects. *Psychological Science* 16(3): 180–3.
Puddifoot, Katherine. 2017. Dissolving the epistemic/ethical dilemma over implicit bias. *Philosophical Explorations* 20(sup1): 73–93.
Saul, Jennifer. 2017. Implicit bias, stereotype threat, and epistemic injustice. In *The Routledge Handbook of Epistemic Injustice*, ed. by Ian James Kidd, José Medina and Gaile Pohlhaus. New York: Routledge, 235–42.
Uhlmann, Eric; and Cohen, Geoffrey. 2006. 'I think it, therefore it's true': effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes* 104: 207–23.
United States Department of Justice Civil Rights Division. 2015. Investigation of the Ferguson Police Department. < https://www.justice.gov/sites/default/files/opa/press-releases/attachments/2015/03/04/ferguson_police_department_report.pdf>