Book Section:
Holroyd, J.D. and Puddifoot, K. (2019) Implicit bias and prejudice. In: Fricker, M., Graham, P., Henderson, D. and Pedersen, N., (eds.) Routledge Handbook of Social Epistemology. Routledge , pp. 313-326. ISBN 9781138858510

https://doi.org/10.4324/9781315717937

This is an Accepted Manuscript of a book chapter published by Routledge in The Routledge Handbook of Social Epistemology on 07/08/2019, available online: http://www.routledge.com/9781138858510

Implicit bias and prejudice

Recent empirical research has substantiated the finding that very many of us harbour implicit biases: fast, automatic, and difficult to control processes that encode stereotypes and evaluative content, and influence how we think and behave. Since it is difficult to be aware of these processes - they have sometimes been referred to as operating 'unconsciously' - we may not know that we harbour them, nor be alert to their influence on our cognition and action. And since they are difficult to control, considerable work is required to prevent their influence. We here focus on the implications of these findings for epistemology. We first look at ways in which implicit biases thwart our knowledge seeking practices (sections 1 & 2). Then we set out putative epistemic benefits of implicit bias, before considering ways in which epistemic practices might be improved (section 3). Finally, we consider the distinctive challenges that the findings about implicit bias pose to us as philosophers, in the context of feminist philosophy in particular (section 4).

1) Good epistemic practice

Let us start by noting various hallmarks of good epistemic practice, as we find them in the epistemology literature, and the ways that implicit biases have been argued to thwart good epistemic practice.

Implicit bias and distortion

Central to the idea of good epistemic practice is the notion of standing in the right sort of relation to the world by tracking the truth in our belief formation and judgement:

> TRUTH-TRACKING: S's belief forming processes are in good epistemic standing if S's belief formation tracks the truth.

For example, suppose at a philosophy conference you form the belief that Jessica is an epistemologist. In order to meet the norm TRUTH-TRACKING you would only believe that Jessica is an epistemologist if it is true that she is; you track the way the world is.

Suppose that we have reason to believe that our belief forming processes lead to systematic distortion of our beliefs and judgements. This would indicate that our belief forming processes do not meet norms such as TRUTH-TRACKING. Jennifer Saul argues that this is precisely what we should conclude from the findings about implicit bias. She claims that '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'* (248). Her concern is that we systematically have our judgements and beliefs distorted by considerations of gender and race and other social identities. Since in the normal course of things, when evaluating evidence or methodology these considerations are not relevant to our enquiries: irrelevant considerations distort our judgements. Let us consider an example of the sort of thing Saul has in mind.

Imagine you are taking part in a psychological study, and your task is to evaluate the importance of traits that you find on the CV in front of you to the role for which individuals are applying (the role of police chief, say). If you are tracking the truth then you will form judgements based on the assessment of the relevance of the traits to the role at issue. Those judgements should be consistent across the CVs you look at: if you think that certain qualifications are very important when possessed by candidate A, but irrelevant when possessed by candidate B, then your assessment of the importance of those qualifications appears not to track the truth about which traits are really relevant to the role.

Yet, this is just the pattern of judgements that Ulhmann and Cohen (2007) found when they asked participants to make such evaluations. When considering the importance of being 'streetwise' to being a police chief, participants tended to judge this as important when a male applicant possessed it, but not when a female applicant did.[1] This led them to make more positive hiring recommendations for the male candidates, irrespective of their differing qualifications for the post. Ulhmann and Cohen conclude that implicit associations between men and police chief roles distorted individuals' clear-eyed evaluation of the importance of qualifying characteristics.

To fully appreciate how people's judgements violate the truth-tracking norm in their responses in this study, consider how they fail to meet some norms which have been taken by various epistemologists to capture what it is for an epistemic agent to track the truth:

> SENSITIVITY: *S*'s belief that *p* is sensitive if and only if, if *p* were false, *S* would not believe that *p* (cf Nozick 1981)
>
> SAFETY: S's belief that p is safe if and only if p could not easily have been false (Williamson 2000, Pritchard 2005).

Suppose that when evaluating the streetwise woman's CV you form the belief: (SW) *being streetwise is* not *important to the role of police chief*. For the sake of argument, suppose it is true that it is not so important to the role. Does this belief track the truth? On one view, we ask: is this belief sensitive, or would you still believe it if it were false? In these scenarios, where the only difference is the gender of the applicant but different responses are made, you seem not to be sensitive to the truth or falsity of the belief - you'd falsely believe it important if a man had that trait. What about SAFETY? Is your belief unsafe: could your true belief easily have been false? It would seem so: even where your belief is true, it could have easily have been false: if you had formed your belief in a situation where the only difference were the gender of the applicant for the job, you may have reached a different belief about the weighting of the trait. If this is right, then it looks like these implicit associations lead people to occupy a poor epistemic situation, forming beliefs that are insensitive or unsafe, even when their beliefs are true.

Implicit bias and perceptual evidence

Consider another marker of good epistemic practice that epistemologists have identified:

> EVIDENCE: S's beliefs have positive epistemic status if (a) the believer has good supporting evidence for those beliefs and (b) the belief is based on that evidence.

We might think that a paradigm case of having good supporting evidence for our beliefs is when we appeal to perceptual evidence under normal circumstances. Suppose you perceive that Jessica is at the conference and form the belief that she is in attendance. Unless you have reason to suppose your perceptual systems to be unreliable under the circumstances in which you perceived Jessica, your belief appears to be evidentially supported. However, some philosophers have worried that implicit biases sometimes pose problems even for our reliance on perception. Jennifer Saul (2013) and Susanna Siegel (2012) have each pointed to studies that indicate that implicit associations lead to distortion of our perceptual judgements. (One way of putting this worry might be that our biases 'cognitively penetrate' our perceptions (Siegel, 2012)).

To see the worry, imagine that you are participating in an experiment in which you are asked to identify a picture of an object that is flashed before your eyes - a weapon, such as a gun,

[1] Ulhmann and Cohen used certain experiences, by which the candidates were ranked, to indicate the extent to which a candidate was 'streetwise': e.g. 'worked in tough neighbourhoods', 'got along with other officers'. This was contrasted with qualifications pertaining to being well educated:'well schooled', 'with administrative experience'.

or a non-harmful object, such as a tool. You might expect the perceptual judgements that you form to be solely determined by the evidence before you (shape of object, features manifest, etc.). However, Keith Payne's (2006) study showed that whether a picture of a black or white male's face was flashed before the picture of the object made a significant difference to individuals' perceptual judgements. If you are primed to think about black persons, you are more likely to perceive that the object is a weapon. This means that the presence of a prime can determine the way the object is perceived. These findings suggest that implicit bias 'affects our very perceptions of the world' (Saul 2013, 246), preventing our perceptions from being fully determined by the relevant evidence available in our environments.

Implicit bias and internalist justification

A further norm of good epistemic practice that some epistemologists advocate is the following:

> ACCESSIBILITY: S's belief that p is justified if and only if S has access to good, undefeated and consciously accessible reasons for believing p.

Those who endorse this norm argue that factors that are under the radar of consciousness are irrelevant to the justificatory status of a belief: they can neither justify nor defeat the justification of a belief.

But now imagine again that you are taking part in the CV study, or Payne's object perception study. Notwithstanding the impact of biases that we have outlined above, it may nonetheless seem, from your own perspective, that you have access to good reasons for your belief - your perceptions of the object, for example. Moreover, the influence of implicit bias on perception may operate under the radar of consciousness: therefore it is irrelevant to the justification of your beliefs according to ACCESSIBILITY. According to this norm, you have justified beliefs, although your beliefs are affected by implicit biases (Puddifoot 2015).

These observations pose difficulties for those who endorse ACCESSIBILITY. There is a strong case for saying that where you form a belief based on bias-inflected perceptions your belief does not conform to the norms of good epistemic practice. But this cannot be captured by the norm ACCESSIBILITY (ibid.). Although firmly committed accessibilists might bite this bullet, this commits them to accepting that their account cannot capture the various ways that implicit biases lead people to deviate from good epistemic practice.

Implicit biases, responsibility and virtues

Some epistemologists have emphasised that good epistemic practice involves forming beliefs responsibly, giving priority to the following norm:

> RESPONSIBILITY: good epistemic practice requires forming beliefs responsibly, in a way that fits with the evidence, coheres with one's previous beliefs, and is based on good reasons (Bonjour 1985, Kornblith 1983).

From what we've already seen – the CV example and the weapons bias case – implicit biases pose difficulties for forming beliefs responsibly: biases prevent us from responding to the evidence that is available to us and from being aware of the real reasons for our beliefs. A subset of epistemologists who place responsibility at the heart of their account of good epistemic practice are virtue responsibilists, who give priority to norms such as the following:

> VIRTUE: good epistemic agents exercise a suite of epistemic virtues, including intellectual carefulness, perseverance, flexibility, open-mindedness, fair-mindedness and insightfulness

> (Zagzebski 1996, 155).
>
> VICE AVOIDANCE: good epistemic agents avoid the exercise of a suite of epistemic vices, including intellectual pride, negligence, conformity, rigidity, prejudice, closed-mindedness and lack of thoroughness (Zagzebski 1996, 152).

Psychologists have emphasised that implicit biases function in the service of efficiency (see Moskowitz and Li 2011): these sorts of automatic processes are useful to us when we're under time pressure, or preoccupied with other tasks. But whilst this arguably might sometimes be useful (see section 3), it is in tension with intellectual virtues such as carefulness, flexibility, and perseverance. The person whose beliefs and judgements are influenced by implicit bias displays epistemic vices like negligence and lack of thoroughness. And since implicit biases often encode stereotypes, they are a hindrance to the achievement of virtues such as open-mindedness, leading to prejudice and closed-minded responses. Note that in some circumstances – e.g. of limited evidence of social equality or in which little is known about implicit bias – agents might demonstrate these vices even whilst doing all they can to live up to their epistemic responsibilities. This may be a case of what Fricker (2016) calls 'no fault' epistemic responsibility: non-culpable epistemic responsibility for a biased response even whilst it is unreasonable to expect her to have avoided it. Implicit biases therefore make both VIRTUE and VICE AVOIDANCE difficult to achieve, even for those who make substantial efforts to meet these norms.

Implicit bias and appropriate trust

Finally, consider the dispositions involved in responding appropriately to testimonial evidence. Elizabeth Fricker (1995) endorses:

> TRUST: good epistemic agents should adopt an appropriately critical stance towards the testimony of others (Fricker 1995).

For example, imagine you tell me that Jessica was at the conference in May. According to TRUST, I should adopt the appropriately critical stance, using markers of credibility to evaluate whether your testimony has evidential value, before trusting (or not) your testimony about Jessica's whereabouts.

As Patricia Hill Collins argued, decisions about 'whom to trust and what to believe' are key to what version of the truth will prevail. And such decisions have often been guided by sexist and racist assumptions (2010, 252). Stereotypes that reduce black women's credibility have been used to disregard black women's testimony and exclude them from the domain of knowers (254, see also 69-97). More recently, Miranda Fricker (2007) has elucidated the way prejudice can inflect trust in her development of the idea of 'testimonial injustice'. Fricker describes how judgements about whether a source of testimony is credible can rely on negative prejudices that track social identity, such as gender or race. Fricker illustrates this with the example from *The Talented Mr Ripley:* Dickie Greenleaf's girlfriend Marge has valuable testimonial evidence about the circumstances of his disappearance, but she is written off as proffering merely 'women's intuition', and as being unreliable due to her distress. Knowledge is lost as she is excluded from the domain of knowers. More generally, TRUST is unlikely to be met in interactions with members of stigmatized groups stereotyped as epistemically inferior in this way: a default position of suspicion which fails to track true markers of credibility may be manifest and an inappropriately critical stance adopted in assessing the testimony of members of such groups.

We might think that so long as we are free from the sorts of prejudices that beset the agents in *Mr Ripley* - outright, or paternalistic, sexism - we are not hindered in our development of appropriate dispositions of trust. But implicit biases might be implicated in our dispositions to trust in complex ways. Some autonomic responses may subtly inflect our interactions in a way that undermines TRUST. For example, Dovidio et al (1997) found differences in the automatic aspects of behaviour of white participants towards black and white interaction partners. In interactions with black partners, white participants' rates of eye-blink, which reflects tension, were higher; and rates of eye-contact, reflecting intimacy and respect, were lower. Interestingly, since the white individuals were not readily aware of these aspects of their behaviour, this also led to divergent impressions of how pleasant the interactions were: the black participants noted the signals of discomfort, whilst the white participants were not aware they were displaying such behaviour (Dovidio et al 2002). The tension that besets, and discrepant impressions of these interactions, Dovidio et al claim, means that these automatic behaviours operate 'in a way that interfere[s] with a foundation of communication and trust that is critical to developing long-term positive intergroup relations' (89). Kristie Dotson (2011) characterises such micro-behaviours as ways in which hearers demonstrate 'testimonial incompetence' and fail to provide the appropriate uptake for the testimony. Having these biases hinders interlocutors in meeting norms of TRUST and important testimonial evidence may be lost (cf Fricker 2007).

Implicit bias, exclusion and epistemic practice

It is worth noting that in various endeavours to address under-representation it has been argued that implicit biases are part of the (complex) explanations for continued marginalisation and exclusion of individuals - in particular women and black and minority ethnicity individuals - from communities of enquirers (Saul 2013). If this is so, then on views according to which diverse communities of enquirers are better positioned to identify errors and biases, implicit biases will be doubly implicated in undermining good epistemic practice: first in distorting the judgement of individuals; secondly in sustaining a homogeneous community of enquirers in which those distortions cannot well be detected and corrected (cf Longino's model of enquiry, 1990).

Implicit bias and scepticism

We might wonder how widespread are the difficulties we have outlined. Saul (2013) argues that these findings mean that we have strong reasons not to trust our own 'cognitive instruments' (to use Hookway's (2010) terminology). Very many of our judgements may be distorted by implicit biases and the chances of detecting and correcting all of these are slim. We may never be sure, Saul argues, that we are forming good judgements - especially if perception itself is subject to distortion, and especially if we can't even detect the occasions on which we are getting things wrong. This sort of 'bias-induced doubt', Saul claims, is more pressing than that induced by more traditional sceptical challenges: we have strong reason to suppose that the challenge is realised (rather than just a theoretical possibility, such as that we are brains in vats); and, the scope of the challenge is broad - it would undermine very many of our judgements and beliefs (rather than some subset of them, as may be the case if we learn we have reason to doubt our probabilistic reasoning). Moreover, it is not clear that this kind of doubt can be overcome by individual exercises of reasoning; social resources may be required to mitigate bias-related doubt (see section 3 below). As such, the phenomenon of implicit bias should lead us to a radical form of scepticism.

One of the premises that provides the basis for the move to pervasive bias-related doubt is that we cannot be sure whether we are affected by biases. Recently, though, contention has emerged regarding whether we can be aware of implicit biases: in part this might depend on the sense of awareness at issue (see Nagel 2014, Holroyd 2014). Recent empirical studies suggest that, at least under certain conditions, individuals can become aware that their behaviours manifest bias (Hahn 2013). If we are able to deploy strategies to reliably track this, the scope of the sceptical challenge could be somewhat limited, by helping us to be aware of those cases in which our judgements are indeed beset by biases. However, the efficacy of those strategies needs further investigation.

Note that this line of scepticism supposes that implicit biases are always epistemically defective - leading us away from epistemic norms such as TRUTH-TRACKING, and hindering our cultivation of good dispositions such as VIRTUE, or TRUST. This stands in stark contrast to another line of reasoning, which has proposed that, whilst morally problematic, implicit biases might nonetheless yield some epistemic benefits.

2) Epistemic benefits of implicit bias?

Here we consider whether (a) there is a case for accepting that implicit biases can sometimes be epistemically beneficial; and (b) there are costs to strategies intended to reduce the influence of implicit bias. Given the claims of section 1, it may seem obvious that implicit biases are damaging and that it is epistemically beneficial rather than costly to reduce the influence of implicit bias. However, these assumptions have been challenged by, in turn, Jennifer Nagel (2012, 2014) and Tamar Szabo Gendler (2011).

Benefits of implicit biases?

Nagel (2014) disputes what she describes as a misinterpretation of empirical evidence that ‘intuitive’ forms of thought - 'type 1', fast, automatic, implicit cognitions - are irrational and likely to lead to errors, in contrast to ‘reflective’ - 'type 2', slow and deliberative - forms of thought, which are taken to be rational.[2]

Nagel defends intuitive reasoning in general - and indeed, there is no reason to suppose that all cognition that is non-reflective is therefore defective - but pertinent to our concerns is her defence of implicit biases in particular. The latter rests on empirical findings suggesting that implicit attitudes can be updated to reflect stimuli presented to the thinker. In an experimental setting, participants were exposed to pairings of black individuals with positive words and images and white individuals with negative words and images. Following exposure to these pairings some implicit associations altered, even in the absence of any change of their explicit attitudes (Olson and Fazio 2006, cited in Nagel 2012). These findings suggest that implicit biases may be as evidence sensitive as explicit attitudes. If this is correct, there is reason to think that implicit biases can be epistemically beneficial; that under some circumstances we are more likely to make accurate judgements that reflect the available evidence if we are influenced by implicit biases than if we are not so influenced.

However, in evaluating Nagel's claims, we should bear in mind the distinction between (i) individuals who hold implicit associations being responsive to evidence and (ii) implicit associations being responsive to evidence. The Olson and Fazio study suggests that implicit biases can be responsive to evidence, and that our associations can change due to exposure to stimuli. But in a way, this is unsurprising, since the associations are held by many psychologists to be the result of associations in our environment to which we are exposed. Moreover, that the associations may be responsive to stimuli does not show that people who hold and are influenced by such implicit associations are more responsive to evidence than they would be if they were not subject to biases. It is consistent with Fazio & Olson's claim that being influenced by implicit bias brings epistemic costs, leading to distorted judgements, and preventing us from forming beliefs based on the evidence. This is so even if the association itself can be altered through strategies such as counter-stereotyping used in the Olson and Fazio study.

[2] Ultimately, Nagel endorses a way of distinguishing intuitive and reflective thinking that identifies the ways each draws on working memory. Each depends on the other (228-231). See also Carruthers 2013.

Epistemic costs of reducing bias?

A distinct challenge is presented by Tamar Szabo Gendler, (2011) who agrees that there are serious epistemic costs to being influenced by implicit bias, but argues that there are also epistemic costs to choosing not to be influenced by implicit bias. To so choose, Gendler argues, is to choose not to be influenced by social category information. For example, to avoid weapons bias of the sort described above, you could ignore social category information about high rates of crime among the Black population of the United States. The result of this choice would be that you would no longer more strongly associate the members of the social category group (e.g. Black people in the US) with the undesirable features (e.g. weapons, or crime) and you could avoid relying on implicit associations that you explicitly repudiate. However, Gendler argues, this choice involves the explicit irrationality of choosing base-rate neglect: the neglect of important and relevant background information (Tversky and Kahneman 1974). Gendler's argument presents a dilemma between two epistemic aims: between avoiding the epistemic costs of base-rate neglect, or avoiding the epistemic costs of the influence of implicit associations (though see Mugg 2013 for a rejection of the claim that there are epistemic costs of being biased to the bearer of the bias).

Another potential dilemma we may face is between ignoring base-rate information - an alleged epistemic cost - and utilising it - an alleged ethical cost, insofar as implicit biases lead to discriminatory differential treatment of members of groups targeted by the biases (see Kelly & Roedder 2008, Brownstein 2015, for articulation of the ethical/epistemic dilemma).

Either way of setting up the dilemma supposes that there are some epistemic costs to not utilising implicit biases, since doing so involves a form of base-rate neglect. Focusing on implicit bias relating Black people with violence or crime, Puddifoot (ms) challenges this claim, arguing that our ordinary ways of using social category information about race and crime that occur in the absence of strategies to prevent implicit bias are substantially different from ideal base-rate use: whilst ideal base-rate information use involves using accurate and relevant background information, our ordinary social category judgements involve inaccurate stereotypes, deploying them where they are irrelevant, and allowing them to distort our perception of, for example, case-specific information about individual crimes, suspects or victims. Accordingly, preventing our judgements being influenced by social category information in order to change or remove implicit biases is not equivalent to ignoring useful information, and so does not involve the same epistemic cost as base-rate neglect.

Another response accepts that there is a dilemma, but argues that we can minimise the costs we face. Madva (2016) argues that if we can limit the influence of implicit associations on judgement and behaviour, we could use information about social reality only where appropriate. For example, implicit associations between black men and crime might influence our thought if we are aiming to understand the social forces that culminate to pressurise young black males into criminal activity without implicit associations (e.g. between black males and crime) then distorting judgement on other occasions. It is an open empirical question, however, whether and under what conditions it is possible to prevent our awareness of background information about social categories from inflecting our cognition with implicit biases. Further, the claim that it is possible to control implicit biases in this way will turn, ultimately, on debates concerning the nature of implicit bias (see Levy 2014, Mandlebaum 2015, Holroyd 2016) and the methods available to mitigate its influence.

3) Improving epistemic practice

Insofar as at least sometimes implicit biases hinder our epistemic practice then we should consider ways in which our epistemic practices might be adapted or transformed in order to avoid the distortions of implicit bias. The discussion above highlights that strategies for combating bias may be not only morally required, but also required if we are to avoid poor epistemic practice. However, there may be competing epistemic considerations that must be weighed in deciding what to do.

Insulating from implicit bias

One way in which we might avoid the epistemic distortions of implicit bias is by insulating our epistemic practices from the possibility of bias: deploying procedures to remove bias-triggering demographic information. For example, anonymised CVs can avoid gender or race biases inflecting the evaluation of the quality of the applicants, and could bring practice into line with norms of TRUTH-TRACKING, or EVIDENCE, and help agents with AVOIDING VICES. However, such anonymisation processes may also involve unwanted epistemic limitations. Suppose we know that women and black or minority ethnicity individuals receive significantly less mentoring in a particular profession. Knowledge of an applicant's race or gender, therefore, could help to contextualise some of the information provided on the CV, and to understand qualifications as achieved despite less mentoring.

De-biasing

Other strategies aim to remove the bias from our cognitions. This includes measures either to 'retrain' associative thinking or affective responses to remove problematic biases; or to train other aspects of cognition to effectively manage and block the manifestation of bias. Studies have suggested a range of surprising measures may be effective, such as: imagining counter-stereotypical exemplars (Dasgupta & Asgari 2004); imagining interactions with individuals from stigmatised racial groups (Crisp et al 2012); retraining approach/avoidance dispositions (Kawakami et al 2007), by way of retraining the associations. Or, to block the influence of bias measures found to have some success include imagining cases in which one has failed to act fairly, thereby activating 'egalitarian goals' (Moskowitz & Li, 2011), or deploying 'implementation intentions - cued cognitive or behavioural responses to environmental stimuli - (Sheeran, Webb & Pepper 2010).

Nagel (2014, 238) has raised concerns about the specific epistemic costs of some of these strategies: implementation intentions, she argues, generate a general loss of accuracy in object identification studies (the weapon/tool studies outlines above). The main general concern for such strategies, however (setting aside for now the issue of whether debiasing involves epistemic loss), are the epistemic difficulties in knowing how effective these debiasing strategies are. On the one hand, some studies have failed to replicate success in mitigating biases; on the other, even if the studies robustly demonstrate bias reduction, it can be difficult to generalise outside of the lab, or to other kinds of bias. Since biases are varied in content and, it seems, in operation, measures which are successful in combating one kind of bias may not be so for others (see Holroyd & Sweetman 2016; Madva & Brownstein ms). For example, bias reduction strategies may aim to reduce certain stereotypical associations, such as those between black people and physical (rather than intellectual) constructs (cf Amodio & Devine 2006). Yet it is unlikely that this same bias is implicated in e.g. perceptions of greater hostility in black facial expressions (Hugenberg & Bodenhausen 2007), or in weapons biases (Payne 2006) – for which different interventions may be required.

Individualistic and interpersonal correctives

Some have argued that individual virtues are an important corrective to combating the distortions of prejudice and implicit biases (see Webber 2016; Fricker 2007). For example, Miranda Fricker has proposed the virtue of 'testimonial justice' as required to avoid unjustly underestimating the credibility of interlocutors. One way in which this sensibility may manifest, Fricker suggests

(2010), is in alertness to cognitive dissonance between judgements of credibility (which may be infected by implicit biases) and the anti-discrimination norms to which one subscribes. Dissonance can provide 'cues for control' and prompt critical assessments of the evidence.

Note that this strategy supposes that individuals can at least sometimes become cognisant of their susceptibility to bias - arresting prejudicial tendencies when it is noticed that they are in operation. Is such awareness possible? There are at least three senses of awareness at issue in the literatures of philosophy and psychology: introspective awareness of the associations; observational awareness of our behaviour being inflected by bias; inferential awareness of our propensity to bias given the empirical findings (Holroyd 2014). Notwithstanding obstacles to awareness such as self-deception or misleading introspective evidence, some recent studies suggest that whilst we are individually poor judges of the extent to which we have or manifest implicit biases, we are nonetheless better at noticing the effects of bias when prompted in interpersonal interactions to reflect on this (Hahn et al 2013).

Even if individuals are able to ascertain that their behaviours manifest bias, it is difficult to detect *to what extent* this is so. It is not the case that whenever we find implicit bias influencing belief, the counterfactual 'She would not believe that but for the bias' will be true: in some cases biases might shore up a belief, make it peculiarly insensitive to revision, but ultimately provide bad epistemic grounds for a belief that has independent epistemic support. So assessing the ways in which biases have affected our beliefs requires careful weighing of evidence with the likely contribution of bias: yet we are not often in an epistemic position to do this (for an example of the difficulties that beset attempts to 'correct' judgements which might be inflected by implicit bias, see Kelly & Roedder's (2008) discussion of grading student papers).

Structural correctives

Haslanger (2015) has suggested that if the correct analysis of injustice is primarily structural, then individual corrective responses are unlikely to effectively target injustice. Moreover, correcting individual implicit biases may be ineffective in the absence of broader social change: few de-biasing strategies have lasting effects, as problematic associations remain in our social environment. Similarly, Elizabeth Anderson has worried that if problems of improper credibility assignment are systemic and structural, then individual virtue as a corrective is unlikely to be sufficient to address the problem: not only because it can be hard to identify when such correctives are needed, and hard for individuals to be constantly vigilant to going wrong; but also because broader structural solutions are needed (2012, 167).

Anderson describes three ways structural change can be beneficial: first, enabling individual corrective measures to work - for example by having institutional procedures that clearly specify explicit grounds for decision-making, sufficient time for making decisions carefully in accordance with those criteria, and accountability for discriminatory outcomes (Anderson 2010, 168). Second, structural changes can instantiate individual virtues at a collective level - an institution may endorse institutional policies that accord with the norm of TRUST or VIRTUE (168-169). Finally, Anderson proposes a more radical kind of structural change: we might hope to foster social and structural arrangements that promulgate 'epistemic democracy: universal participation on terms of equality of all inquirers' (Anderson 2010, 172). This involves ending patterns of informal segregation that structure educational provision, and communities of inquiry.

Avoiding scepticism?

Given that epistemic difficulties beset the remedial strategies themselves - how effective they are, when they are needed, whether they can be successful in the absence of broader structural change - the threat of scepticism may remain until our social and epistemic environment is refigured along the lines of Anderson's 'epistemic democracy' (cf Saul 2013). An alternative response to bias-related scepticism has been proposed by Louise Antony (2016): confronted by the pervasiveness of bias, Antony suggests, we should not retreat to scepticism, but rather adopt a naturalised epistemology of inquiry based evaluation into which biases are a hindrance to good epistemic practice, and which are not. We may be unable to function bias free, but we can weed out the bad ones, and use methods of scientific enquiry to do so.

4) Challenges for philosophers

Finally, philosophical engagement with empirical research on implicit bias raises distinctive epistemic challenges.

Repligate

The discipline of social psychology is itself presently facing a crisis of replication - so called 'repligate'. In a wide-ranging attempt to replicate important findings, only 39 of 100 studies reproduced the original results,[3] prompting critical reflection on the methods of empirical psychology: whether null results should be published, data widely and openly available, methods and analyses pre-registered to avoid selective analysis and so on. We are not suggesting that this provides decisive reason to doubt all of the findings of empirical psychology - especially since some findings have been replicated via robust methods. However, there is reason to adjust our confidence in the outcomes of empirical studies until they have been robustly replicated. And whilst a large number of studies have demonstrated the existence and effects of implicit bias (see Jost 2009), attempts to mitigate bias have been less successfully replicated (Lai et al 2014).

Under-developed conceptual frameworks

A second difficulty is that the psychological research deploys notions that are constructed with experimental purpose, rather than philosophical rigour, in mind. Most simply, psychologists may use terminology ('belief', 'judgement', 'stereotype', 'desire', 'affect') in different ways from philosophers. More problematically, philosophers may inherit modes of discourse - 'implicit bias' itself being a case in point - that is not robustly worked out. Few psychologists agree on what is meant by 'implicit', and some understandings of it - e.g. accessible (though not exclusively) by implicit measure (DeHouwer 2009) - certainly depart from what is commonly meant in philosophical discourse (where this is often conflated with 'unconscious'). These differences are not benign - normative conclusions to do with accountability and remedial obligation may turn on the sense in which biases are 'implicit', and whether this is incompatible with 'awareness' (see Holroyd 2014).

This means that there is scope for what philosophers may sometimes do best - bringing conceptual clarity to a discourse! But whilst the empirical evidence is fast changing, it may be difficult to reach firm conclusions. We take this to speak in favour of more, rather than less, interaction between philosophers and psychologists - to enable fruitful and conceptually clear discourse across disciplinary differences.

Positioning the literature in relation to extant claims

The findings of empirical psychology are often presented as showing to us surprising and troubling aspects of our cognitions, and their implication in perpetuating injustices. But we should ask why these findings are surprising, and what this tells us about our epistemic disposition towards sources of evidence. One putative and unsatisfactory answer would be that empirical

[3] See Open Science Collaboration (2015), DOI: 10.1126/science.aac4716; however, it is worth considering http://alexanderetz.com/2015/08/30/the-bayesian-reproducibility-project/ [accessed 13.01.2016]for useful critical discussion of how to interpret these failures of replication.

psychology reveals a domain of discriminatory behaviour which it was simply impossible to know about prior to the advent of the this research. This answer is unsatisfactory in that in supposing knowledge of such discrimination was inaccessible to us, we fail to engage with other sources of evidence of these patterns of discrimination and ignore testimonial evidence from individuals stigmatised by such biases. Gloria Yamato provides such testimonial evidence, in her writing from 1988, which pre-dates the recent upsurge of interest in implicit bias within psychology and philosophy:

> Unaware/unintentional racism drives usually tranquil white liberals wild when they get called on it, and confirms the suspicions of many people of color who feel that white folks are just plain crazy. [...] With the best of intentions, the best of educations, and the greatest generosity of heart, whites, operating on the misinformation fed to them from day one, will behave in ways that are racist, will perpetuate racism by being "nice" the way we're taught to be nice (Yamato, 2004, p.100).[4]

The point is not that empirical psychology adds nothing to our understanding of patterns of discrimination - of course it has helped us to develop nuanced models of how aspects of our cognition may be implicated in, and perpetuate, various injustices. But we should not suppose that it reveals patterns of discrimination that were invisible prior to this; to do so privileges certain sources of evidence - from communities of academic scientists - over others, namely, the testimony of individuals targeted by those patterns of discrimination and injustice. Indeed, the value of lived experience as a source of knowledge has been emphasised by Patricia Hill Collins. She suggests that within black feminist thought lived experience reveals itself as a more reliable source of understanding than those produced by exclusionary institutions, quoting Hannah Nelson's remark that for her, 'distant statistics are certainly not as important as the actual experience of a sober person' (Nelson, quoted at 257). Yet, this source of knowledge – testimony on the basis of lived experience – has not been valued accordingly in academic communities. The proposal is that we should think carefully about the epistemic status of the empirical literature: for example, we might regard it as providing evidential support for, and thereby vindicating, the lived experiences of individuals who have long reported on this discrimination, whilst also reflecting on which norms dictate that such testimony requires 'vindicating'; rather than as evidence about newly discovered patterns of discrimination. Such a stance highlights the importance of attending to historically marginalised testimonies, and may help to avoid what Kristie Dotson (2012) has named 'contributory injustice'. In a context where various conceptual resources – some marginalised – co-exist, contributory injustice is the maintenance and deployment of structurally prejudiced and exclusionary hermeneutical resources. Whilst being able to articulate experiences via non-mainstream hermeneutical resources, some speakers may fail to receive uptake for her testimony due to the impoverished resources of the hearer.

Feminist epistemology also offers us resources to account for why testimony has been ignored or marginalised. One might see the literature on implicit bias as filling a 'hermeneutical gap' (a notion developed in Fricker, 2007), and offering us conceptual resources that better enable discourse on discrimination and injustice to proceed. Whilst the findings on implicit bias help us to correct that lack of interpretative resources, they also prompt reflection on how those findings are positioned in relation to other sources of evidence, and on the appropriate epistemic dispositions towards both empirical findings and heretofore marginalised narratives of social reality.

[4] It is worth noting that this paper was not available in any of our university libraries, despite being reproduced in many volumes on gender and race - testament itself to what sources of knowledge are deemed important, perhaps.

References

Amodio D.M. & Devine P.G. (2006) Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behaviour *Journal of Personality and Social Psychology* vol.91(4) p.652

Anderson, E. (2012). Epistemic justice as a virtue of social institutions. *Social epistemology*, *26*(2), 163-173.

Antony, L. (2016) Bias: Friend or Foe? Reflections on Saulish Skepticism in Brownstein, M & Saul, J (eds.) *Implicit Bias and Philosophy,* Oxford University Press, p.157-190.

Bonjour, L. (1985). The Structure of Empirical Knowledge (Cambridge, MA: Harvard University Press).

Brownstein, M. (2015). Attributionism and Moral Responsibility for Implicit Bias. *Review of Philosophy and Psychology*, 1-22.

Carruthers, P. (2013). Evolution of working memory. *Proceedings of the National Academy of Sciences*, *110*(Supplement 2), 10371-10378.

Crisp, R. J., & Turner, R. N. (2012). The imagined contact hypothesis. *Advances in experimental social psychology*, *46*, 125-182.

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

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological bulletin*, *135*(3), 347.

Dotson, K. (2012) A Cautionary Tale: On Limiting Epistemic Oppression, *Frontiers: A Journal of Women Studies*, vol.33(1) pp.24-47

Dotson, K. (2011) Tracking Epistemic Violence, Tracking Practices of Silencing, *Hypatia* vol.26(2) pp.236-257.

Dovidio, J. F., Gaertner, S. E., Kawakami, K., & Hodson, G. (2002). Why can't we just get along? Interpersonal biases and interracial distrust. *Cultural Diversity and Ethnic Minority Psychology*, *8*(2), 88.

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of experimental social psychology*, *33*(5), 510-540.

Fricker, E. (1995). Critical Notice: Telling and Trusting: Reductionism and Anti- Reductionism in the Epistemology of Testimony. Mind, 104, 393–411.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing* (p. 7), Oxford: Oxford University Press.

Fricker, M. (2010) Replies to Alcoff, Goldberg, and Hookway, Book Symposium on *Epistemic Injustice: Power and the Ethics of Knowing*, in *Episteme: A Journal of Social Epistemology*, 7(2)

Fricker, M. (2016). Fault and No-fault Responsibility for Implicit Prejudice—A Space for Epistemic Agent-regret. in M.S. Brady & M. Fricker (eds.) *The Epistemic Life of Groups*: *Essays in the epistemology of colectives*, Oxford: OUP.

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, *156*(1), 33-63.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2013). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369.

Haslanger, S. (2015). Distinguished Lecture: Social structure, narrative and explanation. *Canadian Journal of Philosophy*, *45*(1), 1-15.

Hill Collins, Patricia (2010) *Black Feminist Thought: Knowledge, Consciousness and the Politics of Empowerment*, 2nd Edition, Routledge, New York.

Holroyd (2016) What do we want from a model of implicit cognition? *Proceedings of the Aristotelian Society* (2)

Holroyd, J. (2015). Implicit bias, awareness and imperfect cognitions. *Consciousness and cognition*, *33*, 511-523.

Holroyd, J., & Sweetman, J. (2016) The Heterogeneity of Implicit Bias. in Brownstein, M. & Saul, J. *Implicit Bias and Philosophy,* Oxford University Press.

Hookway, C. (2010). Some varieties of epistemic injustice: Reflections on Fricker. *Episteme*, *7*(02), 151-163.

Hugenberg, K. & Bodenhausen, G.V. (2003) Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat, *Psychological Science* vol.14(6) pp.640-643

Kawakami, K., Dovidio, J. F., & Van Kamp, S. (2007). The impact of counterstereotypic training and related correction processes on the application of stereotypes. *Group processes & intergroup relations*, *10*(2), 139-156.

Kelly, D., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass*, *3*(3), 522-540.

Kornblith, H. (1983). Justified Belief and Epistemically Responsible Action. Philosophical Review, XCII, 1, 33-48.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi, S., Sartori, G., Dial, C. M., Sriram, N., Banaji, M. R., & Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*, 1765-1785.

Levy, N. (2014). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Noûs*.

Longino, H. (1990). *Science as Knowledge: Values and Objectivity in Scientific Inquiry,* Princeton, NJ: Princeton University Press.

Madva, A. (2016) Virtue, Social Knowledge, and Implicit Bias, in Brownstein, M & Saul, J. *Implicit Bias and Philosophy,* Oxford University Press.

Madva, A. & Brownstein, M. (ms) The Blurry Boundary between Stereotyping and Evaluation in Implicit Cognition

Mandelbaum, E. (2015). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*.

Moskowitz, G. B., & Li, P. (2011). Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *Journal of Experimental Social Psychology*, *47*(1), 103-116.

Mugg, J. (2013). What are the cognitive costs of racism? A reply to Gendler. *Philosophical studies*, *166*(2), 217-229.

Nagel, J., (2012) "Gendler on alief", *Analysis*, 72(4): 774–788.

Nagel, J. (2014, June). II—Intuition, Reflection, and the Command of Knowledge. In *Aristotelian Society Supplementary Volume* (Vol. 88, No. 1, pp. 219-241).

Nozick, R. (1981). *Philosophical Explanations,* Cambridge, MA: Harvard University Press.

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

Payne, B. K. (2006). Weapon bias split-second decisions and unintended stereotyping. *Current Directions in Psychological Science*, *15*(6), 287-291.

Pritchard, D. (2005). *Epistemic Luck*, Oxford: Oxford University Press.

Puddifoot, K. (2015). Accessibilism and the Challenge from Implicit Bias. *Pacific Philosophical Quarterly*.

Puddifoot, K. (ms). Stereotyping Criminally

Saul, J. (2013). Scepticism and Implicit Bias1. *Disputatio*, *5*(37).

Siegel, S. (2012). Cognitive penetrability and perceptual justification*. *Nous*, *46*(2), 201-222.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

Uhlmann, E. L., & Cohen, G. L. (2007). "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, *104*(2), 207-223.

Webb, T. L., Sheeran, P., & Pepper, J. (2012). Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials. *British Journal of Social Psychology*, *51*(1), 13-32.

Jonathan Webber (2016) Instilling Virtue In Alberto Masala & Jonathan Webber (eds.), *From Personality to Virtue,* Oxford University Press

Williamson, T. (2000). *Knowledge and Its Limits,* Oxford: Oxford University Press.

Yamato, G. (2004). Something about the subject makes it hard to name. in Margaret Anderson and Patricia Hill Collings (eds.) *Race, Class and Gender* 5th Edition, New York Thomson/Wadsworth pp.99-103

Zagzebski (1996). *Virtues of the Mind,* Cambridge: Cambridge University Press.