
Book Section:
Holroyd, J. (2020) Implicit bias and epistemic vice. In: Kidd, I.J., Battaly, H. and Cassam, Q., (eds.) Vice Epistemology. Routledge . ISBN 9781138504431

https://doi.org/10.4324/9781315146058-10

Jules Holroyd, July 2019
j.d.holroyd@sheffield.ac.uk

Bias and Vice

Can implicit biases be properly thought of as epistemic vices? I start by sketching the contours of implicit biases (1), and then turn to the recent claim, from Cassam, that implicit biases are epistemic vices (2). However, I argue that concerns about the stability of implicit biases and their role in producing behavior make for difficulties in establishing that implicit biases of individuals are epistemic vices (3). I then consider a recently developed model which prompts us to consider implicit biases as properties of groups (4). This raises the question of whether implicit biases might constitute collective epistemic vice. I suggest that there is a way to make sense of this claim, but it requires rethinking how we conceptualise collective epistemic vice (5). These re-conceptualisations can be independently motivated. I close by marshalling some considerations in favour of using the terminology of vice to capture these defects of collective epistemic practice (6).

1. What are implicit biases?

Implicit biases are a heterogeneous phenomena. Authors tend to point to various features that implicit biases share: they operate fast and automatically, they may be difficult for the agent to control or be aware of, they may be arational or limited in the extent to which they are guided by the norms that govern other mental states. Biases may differ in the extent to which they manifest each of these characteristics. There are different kinds of biases, that appear to operate in different ways, and may be differently related to individuals' other attitudes, motives, and beliefs. Different biases are related to different kinds of behavior. For example, some biases might affect judgements (how competent another is); others might affect one's manner (e.g. how friendly one is). Exactly how we should characterize implicit biases is contentious, and not a matter that needs to be settled here (see Holroyd 2016 for an overview and critique of various ways of conceiving implicit biases).

The manifestation of implicit biases in behavior is also a complex phenomenon. Certain patterns of implicit biases have been pervasively found in large scale studies that use implicit measures to access biases on which people are unable or unwilling to report: biases against women, black people, minority ethnicity individuals, and other socially stigmatized groups. This is unsurprising, given the interaction between our cognitions and background patterns of social inequality and injustice (see Madva 2016 for discussion of how to understand this interaction). Moreover, that implicit biases are found to be pervasive resonates with reports of discrimination as persistent and pervasive from those who experience it (see e.g. Williams, 2014; Valian, 2005, Sue et al 2008). The pervasiveness with which biases are found has led some to posit implicit biases as an important explanatory factors in understanding persisting patterns of exclusion and discrimination (Greenwald et al 2015). If very many people, even only occasionally, behave in ways that express implicit bias, a pattern of discriminatory judgement and behavior could emerge, with exclusionary consequences.[1] Next, I introduce some particular kinds of associations, and the sorts of behaviours in which we might find these biases expressed.

[1] This pattern may be part of a 'perfect storm' of factors all pointing towards exclusion (see Antony 2012).

Implicit gender bias and judgements of competence or leadership: Various studies indicate gender bias in the evaluation of CVs, whereby women's CVs are judged to demonstrate less competence, or merit lower pay grades, than commensurate CVs of male counterparts (Bertrand et al 2005; Moss-Racusin et al 2012). Studies have also shown that women are more strongly associated with notions to do with the family than with career oriented notions (with which men are associated) (Rudman & Kilianski 2000); and that women are less strongly associated with leadership roles than men (Valian 2005).[2]

At issue here, then, are the associations themselves (between women and family or supporting roles, men and career or leadership roles) and behaviours that appear to be underpinned by implicit biases: judgements of lesser competence, and lesser recognition, or undervaluing, of women's achievements. When women are viewed through the lens of stereotypes, or judged to be less professionally competent, or to have less intellectual acumen or leadership, due to implicit bias, should we think that those who make these judgements display an epistemic vice?

It can be difficult to reach any evaluations or judgements of agents for their performance on implicit measures in laboratory contexts – not least because biases may be visible in these contexts because all else is held fixed (in a way that is rarely the case in 'real world' scenarios). So it will be helpful to have in mind a real world example.

Implicit bias outside the lab: the following scenario is anonymously reported on the 'What is it like to be a woman in philosophy?' blog:

> I was at a bar with three colleagues, each of whom are a) male, b) my friends, and c) self-identified feminists. So there were four philosophers in a bar, at a 3:1 male-to-female ratio. The table was discussing a book that only half in attendance had actually read. Now, I was one of the two folks who had read the book. It should surprise you, then, to learn that for the life of me, I could not get a word in edgewise! 3/4 people were talking, and only 1/3 of those speaking had read the book under discussion, but every freakin' time I tried to speak, I was summarily shut down, talked over, and/or ignored. […] I was disheartened and sad to be treated this way by my friends. I picked up my phone, only to find that it was out of batteries, and tossed it back down on the table, frustratedly. One colleague took notice of my frustration and asked what was the matter, to which I responded rather directly, "Well there is nothing else for me to do at this table, and now my phone is out of batteries." His response? "That sucks. So anyway, how was your weekend with [my partner]?" Shocked and appalled by this totally unnatural segue, I retorted, "We don't have to stop talking about philosophy!" [implying of course: just because you're going to include me, now.] Totally unawares, he sincerely replied, "No! I really

[2] Goff & Kahn (2013) show that in such studies, the paradigm 'woman' that participants have in mind is a white woman. As such, we should be cautious about generalizing these findings to women of colour, who likely face biases that encode the ways in which gender is racialized. Similarly, they urge caution about generalizing studies about associations with black people, which may really hone in on stereotypes about black men. As such there is a lacuna in the research on implicit bias that is only recently starting to be addressed (see Theim et al 2019 on the biases that might target black women in particular).

> wanted to know how your weekend was!" He didn't even realize what he had done. […]
> All three of these guys are my friends, they are self-identified feminists, and they take themselves to be good allies. I'll bet if I told this story back to them in another context, all three of those guys would be appalled. But from the inside, they had no idea what they were doing.[3]

Of course, since implicit measures (or measures of any kind) were not used in this case, we cannot know that implicit biases were at work here. But let us interpret the case on the plausible assumption that this is an instance in which implicit bias is driving the behavior. This seems credible if we accept that the author is accurate in her judgement that her colleagues are friends, feminists and would-be allies. They do not subscribe to the belief that 'women have nothing of value to contribute to philosophical discussions', say; when they are dismissive of her contributions, they seem not to intend to devalue her, nor to realise that this is what they are doing. It is not a stretch to explain such behavior in terms of implicit associations between women and family (rather than career) – especially given their willingness to include her in a discussion about her relationship! – or tendencies to evaluate women as less competent, or to undervalue women's contributions. Such behavior may be influenced by implicit associations such as those manifested in the studies described above. Insofar as such behaviour expresses implicit gender biases, are the philosophers who so behave displaying an epistemic vice?

2. The prima facie case for implicit biases as intellectual vices

In this section, I outline the contours of Cassam's recent claim that implicit biases are epistemic vices.[4]

2.1 Bias impedes knowledge:

On Cassam's 'Obstructivist' account of epistemic vice (2016 and developed in his 2019):

> OBS: an epistemic vice is a blameworthy or otherwise reprehensible character trait, attitude, or way of thinking that systematically obstructs the gaining, keeping, or sharing of knowledge. (2019, 23)

For example, gullibility is an intellectual vice because it is a trait that hinders responsible knowledge acquisition, leading the inquirer to rely on unreliable sources and leap to unsupported conclusions. This analysis of epistemic vice is helpfully expansive, taking in not only traits, but also attitudes or ways of thinking. Cassam is also explicit that cognitive biases, including implicit biases, may also be candidates for epistemic vice. Drawing on the idea (from Banaji & Greenwald, 2016) that implicit biases are 'habits of thought' Cassam argues that implicit biases can be understood as epistemically harmful attitudes (2019, 168-173). The cases on which he focuses are those of weapons biases: the tendency to misperceive items as weapons when primed with black faces (Payne 2006). Such biases are

[3] https://beingawomaninphilosophy.wordpress.com/2014/01/12/insidious-norms/ [posted 2014, accessed April 2019]

[4] It is also plausible that being influenced by bias is related to other epistemic vices: closed-mindedness, dogmatism, epistemic negligence, perhaps. I set aside the interesting task of teasing out the relationship between biases and other vices for another time.

implicated in potentially lethal harms (such as racist patterns of police shootings), as well as epistemic harms (impeding perceptual knowledge). Implicit biases of this sort, and the kind we considered above, seem clearly reprehensible – that is, criticisable – and Cassam makes the case that such biases are blameworthy insofar as agents are what he calls 'revision responsible' for them – that is, if we have the control to weaken or rid ourselves of such biases.[5] Since there is reason to suppose that biases are to some degree malleable, and can be weakened by various forms of self-manipulation (see Holroyd & Kelly 2015), this, together with their obstructive role in inquiry, would suffice on Cassam's account, to make implicit biases epistemic vices. Our examples (from section 1) might be thought to add plausibility to this claim: in experimental studies, gender biases hinder knowledge about the competences and value of women; and knowledge sharing is certainly impeded by the gender biases of the colleagues involved in the 'what is it like…?' example.

However, I want to flag up the following issue now, and will return to (in section 3, below): namely, the requirement, in OBS, that vices *systematically* obstruct knowledge. This claim might pose difficulties for the obstructivist's contention that implicit biases in individuals are epistemic vices.

2.2 Other conceptions of epistemic vice.

Cassam's conception of epistemic vice is not the only one. An alternative view has it that has it that 'vices will be qualities that reliably produce the bad' (Battaly, 2014, 56). For epistemic vices, the bads at issue will, paradigmatically, be false beliefs. On this conception (the reliabilist conception of vice), a trait or mode of thinking – be it hard-wired or acquired – is vicious if it reliably produces bad effects or outcomes. This relation to bad effects is necessary and sufficient for a quality being a vice (57).

Battaly also sketches a 'responsibilist' conception of vice. This view focuses on the blameworthy psychology of the agent: 'bad motives, false conceptions of the good, dispositions to perform bad actions … are required for vice' (2014, 62). This view – the responsibilist view – is driven by the thought that vices are those aspects of our character that are within our control. Another intuition supporting responsibilism is that what matters is that our character expresses what we care about and value (62). I introduce these views to note that Cassam's is not the only conception on which a case might be made for implicit biases as vices. Implicit biases may produce bad epistemic effects (false beliefs), or may be rooted in blameworthy psychologies (see e.g. Holroyd 2012, Holroyd & Kelly 2015, Brownstein 2015).[6] One might think that these alternative conceptions, like Cassam's, could also accommodate the claims that implicit biases are epistemic vices.

However, these accounts commit to the view that implicit biases should *reliably* produce bad effects; or are *stable* traits of the agent. In the next section, I argue these requirements (like that of *systematic* obstruction of knowledge) pose difficulties for establishing that implicit biases, in individuals, are epistemic vices.

[5] But on this account, biases need not be blameworthy – merely criticisable – in order to constitute vices.

[6] Interestingly, in the debate about the blameworthiness of agents for implicit biases, some who have held back from arguing that bias is blameworthy have tried to establish that nonetheless aretaic evaluations of the agent – evaluations that appeal to virtue or vice terms - are nonetheless apt (see e.g. Zheng 2016, Brownstein 2015). These authors (appealingly, I think) detach blameworthiness from the kind of virtue and vice attributions in a way that is starkly at odds with the characterization of vice on the responsibilitst view.

3. The challenges

a. Predictive validity

Does an individual's implicit bias *systematically* obstruct inquiry or *reliably* produce false beliefs (or other epistemic bads)? To systematically impede inquiry, bias need not invariably do so. Cassam rather follows Driver (2001) in requiring rather that the connection between the vice and the bad epistemic outcome be 'non-accidental' (2019, 11). This is intended to rule out cases where luck plays a role: 'to make room for the possibility that epistemic vice can have unexpected effects in particular cases' (12). For example, if an implicit bias occasionally and unexpectedly promoted, rather than obstructed, knowledge that need not undermine its putative status as a vice. But what is meant by 'non-accidental', precisely? In a useful footnote, Cassam asks what we might think of counterfactual scenarios in which something we presently consider a virtue (open-mindedness) '*normally* gets in the way of knowledge' (fn 25 at p.12, my italics)? So our question is whether an individual's implicit biases systematically, that is, normally, or in the usual run of things (without luck or causal deviance), obstruct knowledge.

Recent meta-analyses examining the relationship between individuals' implicit biases and behavioural outcomes are highly pertinent to this issue. Greenwald et al (2015), in their defence of implicit measures, examine the 'predictive validity' of individuals' implicit biases: namely, the extent to which how an individual performs on an implicit measure enables us to predict how they will behave (the correlation between biases and certain behaviours). They point out that the predictive validity of the IAT is what psychologists would call 'low'[7] – that is, an individual's score on an implicit association test (e.g. whether they have associations between women/family and men/career-related notions) does not allow us to predict with confidence how that individual will behave towards women. They take this to be perfectly consistent with their defence of implicit measures – and I return to this shortly – but the point for now is that it problematizes the claim that an individual's implicit bias will be *systematically* obstructive of inquiry, or a *reliable* producer of bad effects. That implicit biases are poor predictors of behaviour suggests that they don't *normally*, in the usual run of things, produce such bad effects. Occasionally they do, but often they do not (producing neutral or non-discriminatory behavior). Note that the bad effects at issue include both behaviours, e.g. how far away an individual might sit from the target of the bias, and judgements e.g. of competence or value. The latter concerns whether knowledge (accurate judgement) is promoted or obstructed: bad *epistemic* effects.

The issue is simply that, whilst implicit biases might be pervasive, they aren't particularly good predictors of whether individuals will behave in discriminatory or knowledge obstructing ways. That is to say, in many instances in which we find an individual harbours an implicit bias, we don't find a strong relationship to such behaviors.[8]

There are various ways in which this issue might be addressed: the first would be to appeal to Cassam's distinction (2019: 58-68) between a character vice – possessed by a person – and a thinking vice – a vicious way of thinking that can be displayed on occasion

[7] But crucially, not as low as argued by the Oswald et al (2013) meta-analysis, whose inclusion critera Greenwald et al critique. Note that their meta-analyses focused on implicit measures of racial attitudes.

[8] Note that this is unequivocally *not* to say that implicit biases might have epistemic benefits (cf. Gendler 2011 for this claim, which I find problematic for the reasons elucidated in Puddifoot 2017 and Saul 2019).

even by those who do not have the related character vice. One could on occasion be e.g. closed-minded, thereby displaying a thinking vice, without being a closed-minded person in general. This could enable us to say that implicit biases are thinking vices: when that mode of thinking is displayed, it obstructs inquiry or produces false beliefs. [9]

However, it is not clear that this deals with the problem. Consider the distinction between the presence of an implicit association in an agent's mental economy (e.g. between women and family oriented notions), and the activation and use of that association in a particular deliberative episode (biased thinking). The meta-analyses don't directly address this issue, but since they concern how individuals perform on an implicit measure (which activates a bias) and their subsequent performance on some behavioural measure, there is reason to believe that they concern episodes of biased thinking. So the meta-analyses should also lead us to conclude that biased *thinking* weakly correlates with behavioural outcomes. That is: whilst an individual may engage in episodes of biased thinking (they might make associations between men and career, and women and family; or might automatically undervalue the qualifications of women), these thoughts may be overridden by other, non-biased aspects of their deliberative processes. The systematic – normal, in the usual run of things – relationship between episodes of biased thinking and biased behavior also faces the challenge from predictive validity.

A second option might be to consider implicit bias as a low-fidelity, rather than hi-fidelity vice (Alfano, 2013, 31-2, discussed in Cassam 2019, 32-34). Hi-fidelity traits, Cassam suggests, require near perfect consistency: one is not generous unless one behaves generously quite consistently. But many ordinary vices, he suggests, are low-fidelity: occasional expression suffices for the vice. One doesn't have to be consistently cruel to be cruel: one episode suffices. Would bias best be construed as a low- or hi-fidelity vice? Are only those who behave in biased ways on a regular basis displaying the vice of bias, or is a one off instance of bias sufficient for someone to qualify as vicious (as is plausibly the case for, e.g. vicious cruelty)?

Whilst Cassam argues that many ordinary vices are low-fidelity, I am inclined to think that bias is akin to closed-mindedness, which Cassam characterizes as a hi-fidelity vice. An individual who is generally open minded, but has a domain in which they display closed-mindedness, seems not to have the vice of closed-mindedness – that domain is one in which they behave in out of character ways. Likewise with bias: an individual who on occasion displays biased thinking need not have the vice of bias – they in this instance behave in a biased way, which is out of character.

Consider this issue of predictive validity in light of the example from the 'What is it like…?' blog. We might reasonably infer that the behavior of the author's colleagues is not routine for them: were that the case, it is perhaps unlikely that the author would describe them as friends, much less feminist allies. Rather, the incident is notable, we can infer, because *even friends and card-carrying feminists* might on occasion manifest implicit biases. As emphasized by researchers on implicit bias, such biases are pervasive and all of us are at risk of, on occasion, manifesting bias in behaviour.[10]

[9] Compare Levy's argument for the conclusion that those who express implicit racial bias are, in some important respect, racist; even if there are other aspects of their character that are not racist, or anti-racist (2016).

[10] Compare the oft-quoted remark from Jesse Jackson: 'There is nothing more painful to me at this stage in my life than to walk down the street and hear footsteps and start thinking about robbery. Then look around and see somebody White and feel relieved' (Remarks at a meeting of Operation

Implicit biases in an individual's mental economy don't appear to produce the relevant (bad) consequences with the required systematicity to establish them as epistemic vices. So it is not clear we can establish the conclusion that implicit bias systematically or reliably impedes inquiry or produces false belief; it is not clear that implicit bias is an epistemic vice.[11]

How do those who defend the explanatory importance of implicit biases in understanding discrimination deal with the issue of predictive validity? Greenwald et al (2015) maintain that implicit biases are significant, despite their low predictive validity, by pointing to the cumulative effects of implicit biases when they are manifested, even just occasionally, by very many people. Using statistical modelling they show that across a large number of people, implicit biases that correlate weakly with individual behaviours could nonetheless manifest in significant behavioural outcomes across the group as a whole. This suggests that we might do better to consider the phenomenon of implicit biases at the level of groups. Before considering this option, let us turn to the other consideration: whether implicit biases might be thought of as *stable* traits.

b. Stability

One might endorse a conception of vice where what matters is that the vices are stable character traits. The issue of the stability of implicit biases has been hotly contested in recent writings. This contention rests on the fact that implicit measures – such as those mentioned in section 1 above – have been found to have low test-retest reliability. That is to say, as Gawronski puts it 'a person's score on an implicit measure today provides limited information about this person's score on the same measure at a later time' (2019, 583). This is not what would be expected if the measures tracked an individuals' stably expressed traits.[12] So, some have concluded that the measures instead access rather more transient states of the agent: what happens to be in mind at a particular time: 'the momentary activation of associations in memory' (Gawronski, 583). If that is the case, then it puts pressure on the idea that implicit biases – as measured by the sorts of tests described in section 1 – are vices. A 'momentary activation' is certainly not a stable trait, which would pose a challenge for accounts according to which epistemic vices should be stable traits.[13]

PUSH in Chicago (27 November 1993). Quoted in "Crime: New Frontier – Jesse Jackson Calls It Top Civil-Rights Issue" by Mary A. Johnson, 29 November 1993, Chicago Sun-Times). The quote is used to illustrate that even those dedicated to anti-racism, and themselves stigmatised by the stereotypes at issue, can on occasion manifest implicit bias. As such, the behaviour of the colleagues in our example is consistent with them being card-carrying feminists (though of course, we rely on the author's description which provides scant information about their commitments, compared to the abundant evidence of Jackson's anti-racist activism).

[11] Note that my claim is not that implicit biases could never be part of an epistemic vice that an individual possesses. In cases where implicit bias props up and is supported by explicit bias, for example, we may well find epistemic vice (and other vices). My claim is simply that implicit bias itself may not meet conditions for epistemic vice.

[12] See also Brownstein et al (2019) for discussion of whether implicit measures access traits (variously construed) or states.

[13] In fact, nor do such transient states seem to qualify as modes of thinking, even. 'Modes of thinking' suggests default assumptions or inference patterns that individuals tend on balance to rely on – not a mere momentary activation captured in laboratory conditions.

A competing interpretation of test-retest reliability findings is to acknowledge that what individuals have in mind on any one occasion is of course dependent on contextual factors, such that it is no surprise to find that across a range of contexts, the extent to which an individual expresses bias on an implicit measure varies. It is after all well known that implicit biases are malleable: they are highly sensitive to features of the context. This has to do both with the person: how tired or distracted they are, on any particular occasion, which affects how susceptible individuals are to implicitly biased modes of thinking. And it takes in features of the situation: with whom one is interacting, what exemplars from different social groups are encountered (stereotypical or counter-stereotypical) (see Dasgupta & Asgari, 2004), the environment in which a person is encountered, what pressures from social norms are exerted, and so on. We store a rather complex set of information, which can include problematic stereotypes and evaluations; which subset of that stored information is activated depends on the context (see Gawronski 2018).

This way of interpreting the findings about test-retest reliability somewhat vindicates the implicit measures: it is not surprising that there is relatively low test-retest reliability. But it still poses a challenge to the idea that implicit biases are stable features of individuals that qualify as character traits in the way the responsibilist requires.[14] Consider again the 'what is it like…?' example. For all we know, the colleagues in this scenario have varying results on implicit measures (this is likely, if they are in keeping with much of the population). And, as noted, to the extent that they display implicit bias here, this seems noteworthy because it is *not* in keeping with the rest of their characters. They may display implicit biases, but they do not appear to evince a stable character trait in doing so. This poses difficulties for any account of vice according to which it is a stable trait.

I have suggested that the recent analyses showing the low predictive validity of implicit biases, and the low test-retest reliability of measures of implicit biases, puts pressure on the idea that implicit biases could constitute epistemic vices in individuals. However, these concerns should not lead us to reduce the extent to which we are concerned about implicit biases. [15] The challenges confront the specific idea that implicit biases in individuals are epistemic vices.[16] But these challenges have also motivated a new way of conceiving of implicit biases, which prompts us to consider the issue of collective epistemic vice. Next, I introduce the new model of implicit biases, and then turn to consider collective epistemic vice.

4. The Bias of Crowds

[14] The idea that individuals' characters are constituted by how individuals react in particular contexts – rather than as context free fixed points – is a familiar and much discussed one (see Brownstein et al in press, for discussion of this issue).

[15] Also for reasons rehearsed in Holroyd & Saul (2019): namely that low predictive validity still gives cause for some concern that biases might, on occasion, manifest; and that the reliability is not markedly worse than other well-established measures; and that the degree of variation on implicit measures is in keeping with a general pattern of expressed biases. One's bias might vary in strength, but less likely in valence.

[16] Denying they are vices is perfectly consistent with thinking they are blameworthy in a range of ways (see Holroyd et al 2017 for an overview of claims about responsibility, blameworthiness, and implicit bias).

Despite the fact that individual's scores on implicit measures are unstable, and vary from one occasion to the next, there is remarkable stability in aggregate levels of implicit biases across groups (Payne et al 2017). This suggests that, whilst individuals' biases are unstable, and individuals' biases do shift, the nature of that individual shift is limited in a way that does not undermine the mean level of bias of a group. Moreover, whilst implicit measures are weak predictors of individual behavioural outcomes, the aggregate implicit biases of a group are more strongly associated with differential outcomes. Payne et al draw on analyses that show that in countries in which the aggregate level of implicit gender bias is higher (in particular, the association with men and STEM subjects), there are greater gender based achievement gaps in science and maths subjects (Nosek et al 2009); in regions with higher implicit racial biases (associating black people with negative notions such as danger or crime) there are greater racial disparities in police shootings (more black people are shot) (Hehman et al 2018). What can explain the stable aggregate levels of implicit bias, and the stronger relationship with disparate outcomes, despite instability and weak predictions generated at the individual level?

Payne et al propose that we should see implicit biases as an attribute of situations or contexts, rather than individuals (2017, 236). By this, I take it that they want to emphasise the contribution of contextual factors to the ways individuals behave, such that patterns of biased behaviour emerge across samples operating within a particular context. Indeed, their spelling out of this claim is that situations, or contexts, encode or contain social stereotypes (we might also appeal to other aspects of a social context, such as scripts, narratives, and aspects of social meaning (cf. Haslanger 2015)). Features of a particular situation affect what is situationally accessible.[17] For example, if a stereotype of women as nurturing carers is prominent, that will affect the extent to which that stereotype is accessible to individuals. Likewise, if the majority of caring roles are in fact occupied by women, or if prominent representations portray women in such roles, this will also affect the extent to which a stereotype is situationally accessible. Since implicit measures record the stereotypes and associations that are accessible, individuals in that situation will display biases (on implicit measures). Indeed, the situationally accessible biases are fairly constant, so if the relevant features of the situation *and all else* were held completely fixed, we could expect that the individual levels of bias expressed would remain fairly constant (there would be good test-retest reliability). But we aren't mere sponges or mirrors of our situations. The extent to which stereotypes are accessible changes for individuals across time and context, depending on who we interact with, what thoughts we have, what our latest interactions or engagements were, how present in mind stereotypes are, and other aspects of our mental lives etc. However, across sample as a whole, the relative constancy of the background situation, and the stereotypes in that context, contribute to a pattern of implicit bias emerging, which is a) more stable, and b) strongly associated with disparate outcomes. I suggest that one way to interpret these claims is that implicit bias is something manifested stably, in a way that affects behavioural outcomes, in collectives or groups.

To speculatively flesh out an example: take the group of academic philosophers in Anglophone institutions. Any individual philosopher, we would expect, would demonstrate varying levels of gender bias on implicit measures. But the situationally accessible

[17] Situational accessibility is contrasted with chronic accessibility (what is available to the agent given their psychological make up), but as the authors note, these two kinds of accessibility will interact (2017, 236).

associations and stereotypes are fairly constant: in addition to the background conditions of gender inequality that prevail in wider society,[18] philosophy is stereotyped as male, much of the canonical literature taught and taken as giving rise to central research questions is by male philosophers, only recently have efforts been made to include more women and scholars of colour in curricula and in research events, and to uncover the contributions of since marginalised philosophers to the canon. Whilst individual measures of implicit bias would vary from day to day (depending on what literature had just been read, with which colleagues one had engaged, what blogs one had read or contributed to), we would expect a fairly stable mean level of bias across a large sample of academic philosophers in Anglophone institutions. And, if in keeping with findings in other contexts, we would expect this to better predict discriminatory outcomes across the profession than individual bias predicts individual behaviours. This example is speculative, since it is modelled on Payne et al's Bias of Crowds way of understanding bias, rather than underpinned by systematically gathered data looking at implicit measures and behavioural outcomes in this context. But of course, it fits with what limited data we do have about gender and under-representation in philosophy[19] and with the fact that plenty of anecdotal evidence points to patterns of (e.g.) gender bias. Many women in philosophy experience some form of gender bias, some of the time;[20] few individuals who (presumably) have implicit gender biases express it all or even much of the time. All that is needed is that many express gender bias some of the time, even just occasionally – as in the 'what is it like…?' example – for deleterious and discriminatory outcomes to take effect. This is explained by the Bias of Crowds model.

On this model, whereby implicit bias is a stable property of groups, and manifests stably in group behaviour, should we think of it as a collective epistemic vice? On the assumption that, at least in the context under discussion, implicit gender biases obstruct inquiry (in the sorts of ways described in our 'what is it like…?' case) and produce the sorts of negative epistemic outcomes associated with exclusion of philosophers who otherwise have much to contribute, I focus on the question of whether it is a *collective* epistemic vice. Much will depend on the conception of collective vice at issue, to which I now turn.

5. Collective vice

The contours of the case – the Bias of Crowds – we have described is as follows: the collective or group at issue is a relatively loosely formed group of individuals: members of a nation, or region, or profession – without any particular institutional structure unifying those individuals. Nonetheless, across those individuals, we find certain patterns of behaviour which produce certain outcomes. These patterns of behaviour are not intentionally co-ordinated. The outcomes are not aimed for. Is it idiosyncratic to think of such cases as instances of collective vice? Loose collectives have been considered candidates for collective virtue or vice before: Slote's (2001) account of group agency extends to societies, broadly construed; Beggs (2003) considers his account of institutional virtue as applicable to the *polis*. It is not uncommon to attribute vices to loosely constituted

[18] As Saul 2013 notes, these wider societal background conditions cannot be the whole of the story, because philosophy is much worse, in terms of gender inclusion, than other subjects in the humanities and most others across academia.

[19] See Holroyd & Saul 2019 for an overview of some of the relevant data on inclusion in philosophy. This draws on data from Beebee & Saul 2011, Norlock 2011, Botts et al 2014 inter alia.

[20] Though as reports on the 'What is it like…?' blog indicate, some of these experiences look to be the result of blatant and explicit sexism.

groups: Medina writes of the epistemic arrogance of the 'powerful and privileged', for example (2013, 31). That the group is loosely constituted need not be an obstacle to seeing the Bias of Crowds as a vice. [21] What matter is whether they meet other conditions for collective vice.

5.1 joint commitment[22]

On one prominent account of collective virtue and vice, what is crucial is that there is a group or collective constituted by individuals operating under a particular practical identity (team member, or participant in some endeavour). Each individual takes on a joint commitment to some virtuous (or vicious) motive, or to some virtuous (or vicious) end that will be pursued by some good (or poor) method (Fricker, 2010, 241, 243).[23] The virtuous members of the night watch each take on a commitment to vigilance, say. Joint commitment, on Fricker's account, involves a practical and cognitive component. Cognitively, the participants each 'take on' a responsibility to do something, and will involve an awareness that one is committing (2016, 245).[24] Practically, this means that reneging on the commitment will be accompanied by, at least, a demand for an explanation.

The Bias of Crowds model obviously won't count as collective vice on this model. It is entirely implausible to suppose that there is a joint commitment to some bad epistemic motive, or bad epistemic end, involved in cases where groups stably manifest implicit biases – that each participant of the loosely connected group has committed to make discriminatory judgements about the value of women philosophers, say, or to ignore contributions, or dismiss lines of argument. Of course, there may be pockets of bad epistemic motives, and there will most likely be bad epistemic outcomes (loss of important knowledge, failures of understanding, fruitful lines of enquiry not pursued). But it is hard to make the case that these are outcomes that members comprising a group commit to pursing, in any meaningful way of understanding that.

Is this the only option, though? Perhaps we need not establish that vicious joint commitments are taken up. Indeed, at some points in Fricker's discussion there is the suggestion that at least in the case of collective epistemic vice (if not virtue), participants need not actively take on a commitment to a bad motive or end; rather, it suffices that they

[21] Note, though, that the sort of groups I have in mind above are unlikely to meet Beggs' (2003) conditions for constituting a collective (solidarity and decision procedures).

[22] An assumption in what follows is that the discussion is premised on an anti-summativist conception of vice – that is, a conception whereby the collective vice is not reducible to vices of the individual. This is precisely what is at stake in discussions of group level implicit biases – the property of the group (stable biases that correlate with disparate outcomes) is precisely what is harder to establish at the individual level. I do not mean to suggest that there is nothing defensible about summativist conceptions, but simply that such accounts will not be the right model for the case in hand. For discussion of summativist and anti-summativist approaches, see Fricker 2010, Lahroodi 2007, Cordell 2017, Byerly & Byerly 2016.

[23] Fricker also notes some reliability condition will also be needed, to ensure the relationship between the motive or method and good outcome.

[24] Though as Fricker emphasizes, it need not involve awareness that one is committing to something *qua* virtue, nor the reliable relationship between that motive or way or proceeding and good outcome.

fail to commit to a good motive or end.[25] In this respect there is an asymmetry between vice and virtue.[26] In her example of the collectively vicious night watchmen – a bunch of slackers who nod off, entertain themselves, and 'in one or another manner signally failing to jointly commit to the end of vigilance' (243), Fricker writes that 'given that vigilance and negligence are exclusive opposites for a night watch, the watch thereby displays the collective vice of negligence' (243). Merely *failing* to commit to some good end can, in some cases, suffice to constitute epistemic vice.

On one reading of Fricker's night watch case is that the failure to commit to a virtue itself signals a vice.[27] This seems to be Fricker's own understanding of the case, and one which applies here, since vigilance and negligence are exclusive opposites, as she puts it. But this is a limitation of her account: insofar as there are virtues in relation to which a failure to commit need not, thereby, signal vice, these cases will not be captured by the joint commitment model. And indeed, there do seem to be such cases. A failure to jointly commit to courage need not signal cowardice; a group that does not jointly commit to generosity need not signal miserly thriftiness. In the context of biases of crowds: we might hope that a group would jointly commit to fairmindedness; but a failure to do so does not, in itself, signify closed-minded prejudice. Such failures might simply signify that a group has other priorities: a commitment to cautious research rather than courage; a commitment to prudential budgeting rather than generosity. Or – particularly in the case of implicit bias – a failure to commit to fair-mindedness might simply signal a failure to realise that any specific commitment on the matter is needed.

There will also be some vices that collectives may manifest without any joint commitment to bad ends or motives, and which are not signaled by failing to commit to the opposite virtue. A collective or group may display the vice of disorganization without having jointly committed to being disorganized. Nor does a failure to commit to good principles of organization signal a commitment to this vice. The UK Government's approach to Brexit negotiations is a good example of this. A group might display the vice of closed-mindedness without having jointly committed to this stance. Nor does a failure to commit to open-mindedness signal a commitment to this vice. The trans-exclusionary organization 'A Woman's Place' is a good example of this. An institution may display the vice of petty bureaucracy without its members having jointly committed to opacity and obstructive modes of operating. Nor does a failure to commit to well justified efficiency signal a commitment to this vice. Various helplines for utilities services exemplify this vice. And, we might contend, a group may display bias without jointly committing to biased ways of thinking. Nor need a failure to commit to fairmindedness signal a commitment to bias.

---

[25] Compare Battaly's concept 2* that requires not that individuals commit to a bad motive, but that they fail to commit to a good motive. On a strong reading, Battaly argues, this is an implausible view (2014, 64).

[26] I have learnt much about these putative asymmetries from discussions with Charlie Crerar. See also Crerar (2018).

[27] An alternative reading would have it that there is tacit joint commitment between the watch members. They are aware that they are each taking on a certain – bad, negligent – way of proceeding, and expect each other to follow suit. This rendering of Fricker's watch case is consistent with her analysis of virtue and vice in terms of joint commitment, but will not capture the Bias of Crowds. Consider the speculative example from the last section: it stretches credulity to suppose that academic philosophers have tacitly committed to ignoring certain contributions, or dismissing and undervaluing lines of argument coming from women.

The joint commitment route to understanding collective virtue and vice, then, does not seem a promising one for capturing biases of crowds as collective epistemic vice. But there seem to be independent reasons for departing from the joint commitment model of collective vice. Some vices that collectives may display – disorganization, closed-mindedness, petty bureaucracy – are ill-captured by the joint commitment model. What other options might there be?

5.2 invisible hand mechanisms & dispositions to behave

A suggestive but under-explored alternative is also present in Fricker's paper: that virtues or vices might emerge by 'invisible hand' mechanisms, whereby the group feature is not reflected at the individual level, but might emerge through and 'be explained by the way in which the individual level features synthesize to create a quite different feature at group level' (239). Fricker's virtue-based example: a jury might be constituted by prejudiced members whose prejudices all cancel each other out, such that the overall judgement reached shows no prejudice or imbalance.[28] However, Fricker doubts such invisible hand accounts are well placed to capture *virtue*, suggesting that the relationship between the supposed virtue and good conduct should not be accidental or a fluke. Rather, 'the good conduct should be performed *because* of the good motive or skill' (239). This non-accidental relationship seems not to be present in cases in which the trait in question (fair-mindedness) emerges because the prejudices happen to cancel each other out. The jury doesn't seem creditworthy for their fair-minded verdict, she suggests, in her view 'the same point applies to vice' (240).

But we have already seen that there could be reason for treating virtue and vice asymmetrically. This may be another instance in which the conditions for virtue and vice are not symmetrical. We could accept Fricker's claim that, in the case of collective virtue, the feature should not emerge accidentally. But in the case of vice, we could maintain that, because negligence is one of the ways that vices can emerge, mere accident of how the individual traits synthesise *can* produce a collective vice.[29]

The general observation that groups could be vicious through negligence seems to open the door to the invisible hand mechanism being one through which genuine collective vices can arise. Suppose a jury is comprised of twelve fair-minded individuals, but they fail to consider the way that, in their group dynamics, these qualities may not be reflected; good norms of group discussion are not established, some members dominate the discussion, assuming that others will speak up if they disagree. Through negligence, the individual features of the group synthesise to produce a poorly functioning collective, that lacks the fairmindedness that each of the constituent members individually possesses; the verdict instead is ill-informed by evidence and manifests closed-minded prejudice. It is 'mere accident' that this feature has emerged, in the sense that the jury did not commit to it, and its emergence is not intentional. That does not undermine the case for such a feature of the group being vicious. The group dynamic will reliably produce epistemically poor

[28] Another example Fricker uses is that of a debating society, the members of which are prejudiced but whose prejudices cancel each other out such that the debate overall shows no prejudice. One might find this example stretches credulity, since a non-prejudiced debate concerns not just the balance of views expressed, but also the contents of what is expressed. For this reason I focus on the jury example (since jury deliberations are not revealed, any prejudiced contents expressed will not be known).

[29] Compare Battaly's remark (2014, 64) that in individuals vices can negligently emerge.

decisions. One might hold, then, that vices can emerge from invisible hand mechanisms, even if virtues cannot.

We can draw on Byerly & Byerly's (2016) account of collective virtue to develop an account of collective vice that can accommodate invisible hand vices. According to their basic account of collective virtue:

> Collective virtue: 'a collective C has virtue V to the extent that C is disposed to behave in ways characteristic of V under appropriate circumstances' (43).[30]

Thus, a jury has the virtue of fairmindedness if it is disposed to behave in ways characteristic of fairmindedness under appropriate circumstances. If the constitution of the jury makes it such that it is so disposed, then it has the virtue of fair-mindedness – whether or not this constitution is 'mere accident', and whether or not the members have taken on any commitments to that end. We can readily apply this analysis to collective vice:

> Collective vice: a collective C has vice V to the extent that C is disposed to behave in ways characteristic of V under appropriate circumstances.

The case for collective vice thus construed may be stronger than for collective virtue, if one is swayed by the idea that the negligent production of vice should be accommodated, even if the accidental production of virtue should not. [31] A jury is closed-minded or prejudiced to the extent that it is disposed to reach prejudiced verdicts. A group is disorganized to the extent that it is disposed to behave in poorly administered ways (failing to have a representative at important meetings; having incoherent policies; uninformed representatives etc). A collective is closed-minded to the extent that it is disposed to behave in closed-minded ways (ignoring important evidence and arguments; selecting only evidence that supports the group's stated aims; question-begging in debates). An institution displays the vice of pettiness if it is disposed to behave in petty bureaucratic ways (opaque and obstructive procedures; unbending and inflexible adherence to protocols; insistence on procedural norms even when irrational). A collective displays the vice of bias to the extent that it is disposed to behave in discriminatory ways (patterns of behavior that disadvantage some demographic; exclusion or devaluing of the contributions of some).[32] Insofar as these examples appear to be cases of vice that collectives or groups manifest, and insofar as they

[30] They offer a more complex formulation of this basic account (at p.43), which makes clear that the virtue can be construed in terms of group-dependent properties that individual members have. However, because I find their argument from multiple realisability convincing (an argument which purports to show that groups can have properties that are not reducible to the individual realisers of those properties), I stick with this more basic formulation. Nothing in the argument turns on this though, so readers are free to substitute the more complex formulation from Byerly & Byerly should they see fit.

[31] I want to remain agnostic on what we should say about invisible hand mechanisms producing virtue. My main point is that whatever we say about virtue, a case can be made that collective vices can emerge through these invisible hand mechanisms, negligence being one of the key ways in which they can do so.

[32] Note that whilst the emergence of the group level property is not intentional, in the case of bias it is not 'mere accident'; social structures of racism and sexism are effective engineers of these group level properties.

are manifested without joint commitment to some bad motive or end, they can nonetheless be accommodated by the dispositional account of collective vice.[33]

This account has some advantages over the joint commitment account: as Byerly & Byerly point out, a group may commit to virtue without, in fact, being disposed to behave in virtuous ways. Commitments count for little if they are empty. Invisible hand cases also point to the importance of dispositions to behave, rather than commitment. If vices can emerge in collectives, and these are manifested in the dispositions to behave of the collective, then again, the importance of commitment – at least for some vices – is undermined. If one thinks an account should capture invisible hand cases (of vice, if not of virtue), this will also be an advantage of the dispositional, over the joint commitment, analysis. There is independent motivation, then, for moving away from the joint commitment account as providing necessary conditions for collective vice. Some collective vices might be instantiated through joint commitments to bad ends. But others may emerge through invisible hand mechanisms and manifest in the collective's dispositions to behave in ways characteristic of that vice.

Where does this leave us in thinking about implicit bias? This way of making sense of collective vice is particularly helpful for thinking about the Bias of Crowds model. On this analysis a group has the vice of bias when it is disposed to behave in ways characteristic of bias under appropriate circumstances. For example, the collective of academic philosophers in Anglophone institutions would have the disposition to gender bias to the extent that the collective is disposed to behave in ways characteristic of gender bias (undervaluing women's contributions, practices that exclude women from participation in research events, failures to represent women's contributions to the discipline on curricula, and so on).[34] We could also appeal to further evidence (where it is available) of such dispositions: a stable mean level of bias found across a group would be strong evidence that the collective has the relevant disposition. And, such dispositions could be evidenced where there is a strong relationship with disparate outcomes for different demographics within that group.

As noted, the collectives at issue in the Bias of Crowds model are large and loosely connected ones: nations, populations across certain regions. The extent to which we consider these samples as collectives will rest on questions in social metaphysics that cannot be settled here. But there does not seem to be any obvious reason for which we should not treat such large samples of individuals as collectives, if we find stable propensities to behave across such populations.

This view of collective vice will face an objection recently advanced by Cordell (2017): that what I have identified is a feature of a collective, but does not amount to a substantive vice.[35] This is for two reasons: first, he argues that if a feature is to be diagnosed

[33] We might ultimately come to quite different views regarding the vices of institutions and groups or collectives on a number of matters, such as their collective responsibility and blameworthiness, as well as forward-looking responsibilities for correcting vice. These issues, which would have to address the hierarchical structures and power dynamics involved in each, are beyond the scope of this paper.

[34] Beggs suggests that 'practice' might be considered the group analogue to individual disposition (2003, 51). Practice on his account is understood as 'the social grammars (the types) that an individual agent's actions manifest (the tokens) (466).

[35] He advances another line of objection, targeted at Fricker's joint commitment account: that she has not provided an account of an *irreducibly* collective virtue – rather, he argues, the virtues can be reduced to the commitments of individuals in their group oriented roles. Since I think there are

as a substantive virtue or vice, then it must be something that the agent (the collective) can reflect on as something to be cultivated or eliminated from their functioning. But collectives of this sort (he argues) lack the requisite processes of reflection. Second, Cordell suggests that one could avoid this first concern by being purely instrumental about virtues or vices: whatever feature produces good or bad effects (irrespective of any mechanism for reflection on these features) is a virtue or vice of the collective. But this instrumentalist picture is not well suited to capture the extent to which the collective is an agent: the group has a feature, but it is not a feature produced by the *agent*.

These objections may have some promise when directed towards an account of collective virtue. Perhaps for a trait to be genuinely credit-worthy it does have to be intentionally produced – perhaps via mechanisms of reflection – by the collective agent. However, I see no reason to accept these claims with respect to collective vices. As we have seen, vices of collectives could result from negligence, and so by their nature will not be the result of intentional production, or the fruits of a reflective mechanism that has decided to cultivate a particular feature. My view is that it would be an excessively restrictive view of collective vice to insist that they cannot be produced by negligence.

In sum: I have argued that we have good reasons to reject the claim that collective vice requires joint commitment to some bad end or motive, and that a case can be made for vices that emerge – through negligence – via 'invisible hand' mechanisms. This seems true in at least some cases for vices such as closed-mindedness, prejudice, disorganization, or pettiness – and, in particular, bias. This can be captured by an account of collective vice, drawing on Byerly & Byerly, that focuses on the disposition to behave in ways characteristic of vice. Where these dispositions or propensities affect knowledge seeking activities, then, they can be properly described as collective epistemic vice. Implicit biases, when understood on the Bias of Crowds model, are contenders for collective epistemic vice.

6. Vice charging, individual and collective

I have suggested that there are obstacles to determining that implicit biases are epistemic vices in the individual case. But I argued that we should think collective vices can be captured on the 'disposition to behave' analysis, and can emerge without joint commitments. On that analysis, we can claim that implicit biases manifested by groups – the Biases of Crowds – are collective vices. Where patterns of implicit bias across groups serve to obstruct knowledge-seeking or produce bad epistemic effects they will be collective epistemic vices.

But why should we *want* to be able to make such a claim? What is gained by being able to diagnose biases as vicious? What is the advantage of being able to call out collectives as vicious?

As Kidd argues, charging an agent with vice should serve an ameliorative function (2016, 192); the aim should be to do so in constructive spirit, with a view to improving the character or conduct of others.[36] There are good reasons to suppose that characterizing

---

other reasons to depart from Fricker's joint commitment account, I set aside this concern here. It is clear that the feature of the group with which I am concerned – bias – is irreducible to members of the collective, given the considerations raised in section 3.

[36] Kidd also notes that vice charges should 'build in a suitably complex account of agential epistemic responsibility' (2016, 194) and in particular one that is sensitive to the aetiology of the vice. As noted in footnote 6, I find attractive a view according to which vice attribution does not depend on

patterns of behavior of groups and collectives as vices can serve an ameliorative function: first, doing so identifies a systematic defect in the conduct of the collective – a defect which many of the individuals comprising the collective would find reprehensible. This is particularly so in instances where the defect has emerged from invisible hand mechanisms, and where no individuals have committed to bringing about the conduct or outcomes that have emerged. Second, vice charging in the case of collectives can prompt members of the collective to reflect on how they sustain certain patterns of behavior, albeit unintentionally. It can draw the attention of individuals to ways in which they, with others, are complicit in problematic patterns of behavior and outcome, despite their individual subscription to good values, or despite good individual intentions. Thirdly, drawing attention to individual's roles in perpetrating collective vices, in this way, might be a particularly good way of motivating change. Fourth, this is particularly so because it prompts members of the collective to focus not just on what they do, qua individual, but also on the structures, norms and practices that enable these vices to be enacted at the level of the collective. Finally, it focuses attention on what collective measures are needed to avoid these problematic dispositions, and highlights the importance of *collective*, rather than individual, virtue in addressing these issues.[37] Seeing the Bias of Crowds as a collective epistemic vice, then, may serve an important ameliorative function in addressing the problematic patterns of bias in which we are implicated.

Wordcount: 8,063

References.

Alfano, M. (2013). *Character as moral fiction*. Cambridge University Press.

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

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

Banaji, M. R., & Greenwald, A. G. (2016). *Blindspot: Hidden biases of good people*. Bantam.

Battaly, H. (2014). Varieties of epistemic vice. *The ethics of belief*, 51-76.

Battaly, H. (2015). Epistemic virtue and vice: Reliabilism, responsibilism, and personalism. In *Moral and intellectual virtues in western and chinese philosophy* (pp. 109-130). Routledge.

establishing blameworthiness. Of course, there will be many interesting and complex questions to address regarding collective responsibility or blame for implicit bias. In particular, it will be important in this context to be sensitive to the power dynamics within the group, especially when it comes to forward-looking responsibility: whose responsibility it is to undertake, or lead the way in taking, corrective steps.

[37] See Anderson (2012) for concerns that a focus on individual virtue is an insufficient corrective for addressing implicit biases.

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

Beggs, D. (2003) ‘The Idea of Group Moral Virtue’, *Journal of Social Philosophy*, 34.3, 457–74 <https://doi.org/10.1111/1467-9833.00194

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit discrimination. *American Economic Review*, *95*(2), 94-98.

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

Brownstein, M. (2016). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology*, *7*(4), 765-786.

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

Brownstein, M., Madva, A., & Gawronski, B. (2019). What do implicit measures measure?. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1501.

Byerly,T. R. and Byerly, M., (2016) Collective Virtue’, Journal of Value Inquiry, 50.1, 33–50 <https://doi.org/10.1007/s10790-015-9484-y

Cassam, Q. (2016). Vice epistemology. *The Monist*, *99*(2), 159-180.

Cassam, Q. (2019). *Vices of the Mind*, Oxford University Press.

Cordell, S. (2017), ‘Group Virtues: No Great Leap Forward with Collectivism’, *Res Publica*, 23, 43–59 <https://doi.org/10.1007/s11158-015-9317-7

Crerar, C. (2018). Motivational Approaches to Intellectual Vice. *Australasian Journal of Philosophy*, *96*(4), 753-766.

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.

Fricker, M. (2010). 10. Can There Be Institutional Virtues?. *Oxford Studies in Epistemology*, *3*, 235-253.

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*, 1745691619826015.

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

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

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

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

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social psychological and personality science*, *9*(4), 393-401.

Holroyd, J. (2016, July). VIII—What Do We Want from a Model of Implicit Cognition?. In *Proceedings of the Aristotelian Society* (Vol. 116, No. 2, pp. 153-179). Oxford University Press.

Holroyd, J. D., & Kelly, D. (2016). Implicit bias, character and control. In Masala, A., & Webber, J. eds. *From personality to virtue: essays on the philosophy of character*. Oxford University Press.

Holroyd, J., Scaife, R., & Stafford, T. (2017). Responsibility for implicit bias. *Philosophy Compass*, *12*(3), e12410.

Holroyd, J. & Saul, J. (2019) Implicit Bias Research and Reform Efforts in Philosophy: A Defence, *Philosophical Topics*, vol.48 no.2

Kidd, I. J. (2016). Charging others with epistemic vice. *The Monist*, *99*(2), 181-197.

Lahroodi, R. (2007) 'Collective Epistemic Virtues', *Social Epistemology*, 21.3 281–97 <https://doi.org/10.1080/02691720701674122>;

Levy, N. (2017). Am I a racist? Implicit bias and the ascription of racism. *The Philosophical Quarterly*, *67*(268), 534-551.

Machery, E. (2016): 'De-Freuding Implicit Attitudes'. In Michael Brownstein & Jennifer Saul (eds.) *Philosophy and Implicit Bias* Oxford University Press

Madva, A. (2016). A plea for anti-anti-individualism: How oversimple psychology misleads social policy. *Ergo, an Open Access Journal of Philosophy*, *3*.

Medina, J. (2013). *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and the social imagination*. Oxford University Press.

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

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

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., ... & Kesebir, S. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593-10597.

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

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

Puddifoot, K. (2017). Stereotyping: the multifactorial view. *philosophical topics*, *45*(1), 137-156.

Rudman, L. A., & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and social psychology bulletin*, *26*(11), 1315-1328.

Saul, J. (2013). Implicit bias, stereotype threat, and women in philosophy. *Women in philosophy: What needs to change*, 39-60.

Saul, J. (2018). (How) Should We Tell Implicit Bias Stories?. *Disputatio*, *10*(50), 217-244.

Sue, Derald Wing, Christina M. Capodilupo, and Aisha Holder. "Racial microaggressions in the life experience of Black Americans." *Professional Psychology: Research and Practice* 39, no. 3 (2008): 329.

Slote, M. (2001) *Morals from Motives* Oxford University Press

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

Valian, V. (2005). Beyond gender schemas: Improving the advancement of women in academia. *Hypatia*, *20*(3), 198-213.

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

Insidious Norms, post at What is it Like to be a Woman in Philosophy?https://beingawomaninphilosophy.wordpress.com/2014/01/12/insidious-norms/ [posted 2014, accessed April 2019]

Zheng, R. (2016). Attributability, accountability and implicit attitudes. In Brownstein, & Saul (Eds.), *Implicit bias and philosophy* (Vol. **2**) (pp. 62– 89). Oxford, UK: Oxford University Press.