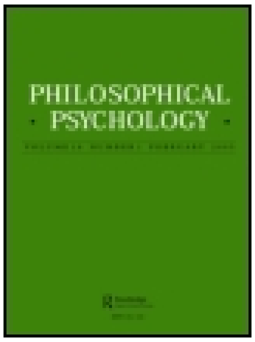

Philosophical Psychology

Routledge
Taylor & Francis Group

ISSN: 0951-5089 (Print) 1465-394X (Online) Journal homepage: http://www.tandfonline.com/loi/cphp20

Implicit racial bias and epistemic pessimism

Charles Lassiter & Nathan Ballantyne

To cite this article: Charles Lassiter & Nathan Ballantyne (2017): Implicit racial bias and epistemic pessimism, Philosophical Psychology, DOI: 10.1080/09515089.2016.1265103

To link to this article: http://dx.doi.org/10.1080/09515089.2016.1265103

Published online: 12 Jan 2017.

PHILOSOPHICAL PSYCHOLOGY, 2017
http://dx.doi.org/10.1080/09515089.2016.1265103

Implicit racial bias and epistemic pessimism

Charles Lassiter[a] and Nathan Ballantyne[b]

[a]Department of Philosophy, Gonzaga University, Spokane, WA, USA; [b]Department of Philosophy, Fordham University, Bronx, NY, USA

ABSTRACT
Implicit bias results from living in a society structured by race. Tamar Gendler has drawn attention to several epistemic costs of implicit bias and concludes that paying some costs is unavoidable. In this paper, we reconstruct Gendler's argument and argue that the epistemic costs she highlights can be avoided. Though epistemic agents encode discriminatory information from the environment, not all encoded information is activated. Agents can construct local epistemic environments that do not activate biasing representations, effectively avoiding the consequences of activation. We conclude that changing our local environments provides a way to avoid paying implicit bias's epistemic costs.

ARTICLE HISTORY
Received 5 April 2016
Accepted 27 October 2016

KEYWORDS
Dual systems; epistemic pessimism; implicit bias; prejudice; epistemic dilemmas

Introduction

Living in a society structured by race obviously presents moral challenges. Tamar Gendler (2011) has argued that it also brings a distinctively epistemic challenge: members of a racially structured society cannot be fully epistemically rational. By living in such a society, we are forced to pay some epistemic costs, even if we happen to explicitly disavow the racial categories and embrace an egalitarian, non-racist perspective.

Gendler's argument is our focus here. Gendler presents it as a straightforward dilemma, but we argue that is misleading. By treating her argument as an either/or, it is easy to overlook crucial details that reveal how it's possible to avoid the epistemic costs she identifies.

We will clarify Gendler's argumentative strategy and suggest a different way to think about the challenges to which she has drawn attention. We will explain why it isn't impossible for members of a racially structured society to be fully rational. Gendler is correct to draw attention to distinctively epistemic problems raised by racial categories, we think, but these problems do not support her pessimistic conclusion.[1]

CONTACT Charles Lassiter ✉ lassiter@gonzaga.edu, charles.lassiter@gmail.com

1. Gendler's dilemma(s)

Gendler presents two forms of her dilemma. Her reasoning is intricate, and so we'll need to reconstruct carefully what she says. The first dilemma runs as follows, though by her admission it is not a "completely accurate" summary (2011, p. 37):

> **G1.** An epistemic agent living in a society structured by racial categories which she disavows must *either* fail to encode the relevant base-rates and cultural background information *or* encode the relevant base-rates and cultural background information. If the agent fails to encode the relevant base-rates, then she pays the epistemic cost of base-rate neglect. If she encodes the relevant base-rates, then she experiences cognitive depletion by regulating the chains of associations that are activated in virtue of the encoded base-rates, which is itself an epistemic cost (see 2011, p. 37).

Here is a second, slightly different dilemma that also appears in her paper:

> **G2.** An epistemic agent living in a society structured by racial categories which she disavows must *either* fail to encode information about racial inequality *or* encode it. If she doesn't encode information about racial inequality, then she is explicitly irrational through base-rate neglect. If she encodes information about racial inequality that she reflectively rejects, then she is implicitly irrational (see 2011, p. 57).

In both G1 and G2, Gendler takes her dilemma to support a kind of epistemic pessimism: "Racially-based inequities – and the psychological processes by which we inevitably encode them – carry not merely moral, but also epistemic, costs. And they carry them regardless of what we believe" (2011, p. 57).

What are the epistemic costs Gendler has in mind? She focuses on a number of psychological effects that automatically and unconsciously compromise reasoning abilities, including base rate neglect and stereotype threat. We'll describe both.[2]

Gendler draws on a study by Tetlock, Kristel, Elson, Green, and Lerner (2000) to illustrate base rate neglect. Subjects were asked to set insurance premiums for neighborhoods in Columbus, Ohio. Experimenters told participants that higher rates had to be assigned to higher risk neighborhoods in order for the insurance company to make a profit. When race went unmentioned, subjects assigned higher premiums for high-risk neighborhoods. If subjects learned that high-risk neighborhoods were predominantly Black, then the base rates became off-limits, especially for political liberals. Subjects who ignored base rates defended their position by appeal to moral considerations: they had no desire to add to the oppression of a marginalized group. While the sentiment is morally laudable, ignoring base rates about neighborhood crime is an epistemically poor move when setting insurance premiums.

Gendler also draws upon research on stereotype threat. Research has revealed that simply activating a subject's thoughts about his or her membership in a group that is typically associated with poor epistemic performance in some domain is enough to decrease the subject's performance. For instance, research by Shih, Pittinsky, and Ambady (1999) is concerned with stereotypes about the mathematical abilities of women (stereotyped as bad) and Asians (stereotyped as good). They asked college-aged, Asian-American women to complete a questionnaire that included questions about single-sex vs. co-ed housing, which made them think (however fleetingly) about being female.[3] Women who completed the questionnaire performed worse on a subsequent math test than women who did not.[4] The evidence suggests that merely prompting women to think about gender was enough to invoke the stereotype that women are bad at math and thereby cause a decrease in performance.[5] Put roughly, reminding women that they are women takes an epistemic toll. Stereotype threat is present in cases involving race, too, and Gendler's idea is that such effects are an epistemic cost produced by racial inequality.

We will say more about how G1 and G2 reflect Gendler's purposes, but first notice that both G1 and G2 share in common an important idea. Living in a racially structured society raises a dilemma of the following form:

D1. Either we encode information about racial inequality or we do not encode it.[6]

G1 and G2 both assume D1, and without D1 Gendler can't reach the conclusion that there are inevitable epistemic costs to pay. But D1 slides over important complexities.

Here is the trouble. Often, when agents encode information about the world, they represent features of the world. When smelling burnt toast, for example, Smith represents the content *there is burnt toast*. Activating encoded information typically happens in one of two ways. First, sensory stimuli can activate encoded information: smelling the burnt toast tokens a burnt-toast representation. Second, an activated representation R1 activates another representation R2 with which R1 is associated. For example, when Smith smells burnt toast, she may be reminded of a specific person or event, or perhaps experience a particular sensation. But if the smell of burnt toast is not perceived, then neither is the memory tokened nor the sensation experienced. The point is that not all encoded information gets *activated*.

This same lesson holds for information about racial minorities: encoding such information does not by itself generate epistemic or cognitive consequences. What generates epistemic consequences is activating the relevant information-bearing representations. Agents might encode the culturally salient (and false) information that Blacks are typically criminals, but that information does not affect agents' judgments if the representation is not activated. The representation has been encoded, but that alone is not enough to compromise the agents' epistemic or cognitive scruples. Dormant information is no immediate epistemic threat.

D1 alone will not prove helpful, then, in trying to determine the epistemic costs of living in a racially structured society. That's because D1 fails to acknowledge that the *activation* of biased representations is important. Again, encoding or not encoding biased representations is not what matters here – whether or not those representations get activated is critical. In place of D1, we offer the following claim:

D2. If we encode information about racial inequality, then that information is either activated or it is not.[7]

We submit that D2 should be a central component of Gendler's argument. (Importantly, there are different ways for activated information to be represented in an agent. We return to this matter shortly.)

For now, consider again G1 and G2. Gendler seems to take G1 and G2 as roughly similar summaries of her dilemma. But G1 and G2 are different. In G1, the proposed epistemic costs are either (1) base rate neglect or (2) regulation of chains of representations. But in G2, the epistemic costs are either (3) explicit irrationality or (4) implicit irrationality. However, the cost of (1) base rate neglect is not equivalent to (3) explicit irrationality and the cost of (2) association regulation is not equivalent to (4) implicit irrationality.

After untangling these threads, we will explain why G1 and G2 are crucially different. First of all, consider costs (1) and (3) that come from encoding racial categories. The epistemic cost of base rate neglect can plausibly occur inside or outside an agent's conscious awareness as Tetlock and colleagues note (2000, p. 864).[8] The cost of *explicit* irrationality, however, falls within an agent's conscious awareness by definition. It seems odd to identify an obviously explicit effect with one, such as base rate neglect, that is plausibly explicit *or* implicit.

Second, there are costs, (2) and (4), that come from failing to encode racial categories. The cost of (4), implicit irrationality, occurs outside an agent's conscious awareness – that is the point of using "implicit" to describe the agent's irrationality. But the cost of (2), association regulation, requires that the associations occur within an agent's conscious awareness – how could a thinker pay the cost of trying to regulate chains of associations unless she is aware of them? Implicit irrationality is a cost that obtains without an agent's awareness, and association regulation is a cost that depends on an agent's awareness.

This strongly suggests that G1 and G2 are not different ways of saying the same thing. But notice a further point. Though Gendler uses information encoding as the dividing line between potential epistemic costs, she seems to think some of the epistemic costs are generated by fast, automatic System 1 processes, while other costs are generated by slow, deliberative System 2 processes.[9]

By our reading, then, Gendler does not pose a single dilemma: G1 and G2 represent two distinct strands in her thinking. Here is what we mean. G1 and G2 can't be underwritten by a single either/or claim about information-encoding, as we find in D1. A taxonomy identifying epistemic costs as products of System 1 or System 2 activities offers a more coherent way to capture Gendler's thinking. And we have also noted that D1 glosses over the psychologically important issue of stored but unactivated representations. We suggest that D2 is needed to make sense of the dilemmas G1 and G2 since what matters most is *not* whether the information is encoded but rather whether the relevant representations are *activated*. And the relevant representations can be located in either the fast, automatic System 1 or the slow, deliberative System 2.

To sum up: Gendler has identified two types of problematic consequences of living in a racially structured society. There's what happens within an agent's awareness and there's what happens outside an agent's awareness. If the relevant representations are activated, then the epistemic consequences differ according to whether the representations are located in System 1 or System 2.

None of this means there is not an argument in the neighborhood of the considerations Gendler has noted for the conclusion that we must pay epistemic costs. In fact, her discussion strongly suggests the following argument, we think:

(1) An epistemic agent living in a society structured by racial categories which she disavows must *either* fail to encode the relevant base rates and cultural background information *or* encode the relevant base rates and cultural background information. (Note: this is D1.)
(2) If the agent fails to encode the relevant base rates, then she pays the epistemic cost of implicit base rate neglect (G1, cost of failing to encode).
(3) If the agent encodes information about racial inequality, then that information is either activated or it is not. (Note: this is D2.)
(4) If the information is activated, then the kinds of epistemic costs the agent must pay depend on whether the information is encoded in System 1 or System 2.
(5) If the activated information is encoded in System 1, then the agent pays the epistemic cost of encoding information she reflectively rejects (G2, cost of encoding).
(6) If the activated information is encoded in System 2, then the agent pays the epistemic cost of cognitive depletion in association regulation (G1, cost of encoding) *or* explicit base rate neglect (G2, cost of failing to encode).[10]
(7) If the information is not activated, then there are no epistemic costs.

Our reconstruction of Gendler's argument helps to identify various consequences of living in a racially structured society – see steps 1 through 6 – but it taxonomizes those costs in terms of both Systems 1 and 2. Steps 2 and 6 express the costs Gendler identifies in G1: (implicit) base rate neglect through failure to encode and cognitive depletion through success in encoding. And steps 5 and 6 feature the costs identified in G2: explicit base rate neglect through failing to encode and representing information that would be rationally rejected were encoding successful. Representing information that would be rationally rejected is a significant contributor to unconscious discriminatory behaviors, including stereotype threat and implicitly biased judgments. By reconstructing the argument in terms of Systems 1 and 2 costs and highlighting the importance of representation activation, we see how encoding racial categories can lead to the epistemic costs that validate Gendler's epistemic pessimism (see steps 2, 5, and 6). But we also see how such costs can be avoided (see step 7).

We have so far offered a detailed reconstruction of Gendler's reasoning behind her epistemic pessimism. In the sections that follow, we'll explore Gendler's argument by addressing the following issues:

- System 1 and System 2 activities are mutually influencing. The representations activated in one affect patterns of activation in the other. We discuss this in section 2.
- The effects of living in a racially structured society affect both the fast, automatic System 1 and the slow, deliberative System 2. The consequences for these different systems are not the same, however. That is partly because it is not clear whether the contents of the representations are the same for both System 1 and System 2. We take this up in section 3.
- Evidence suggests that agents can encode the relevant base rates and yet minimize the epistemic consequences of living in a racially structured society by constructing environments that fail to *activate* those patterns of representations leading to discriminatory behaviors. Given step 7 of our reconstruction of Gendler's argument, this evidence suggests a means to avoid epistemic pessimism. We turn to this in section 4.

2. Systems 1 and 2

We will now briefly consider the operation of System 1 and System 2.[11] Saying exactly how they work is difficult.[12] Our purposes are more modest: we only need to highlight that the activities of System 1 and System 2 are mutually influencing. This will be a key point for avoiding Gendler's epistemic pessimism because it opens up the possibility of encoding bias in System 1 while minimizing the bias's effect on System 2 functioning through failure to activate the biased System 1 representations.

System 1 works rapidly and automatically. When we experience fear while looking over the edge of a cliff that feeling is a product of System 1 activities: we look over the cliff and System 1 tokens a representation whose content includes "Cliff … danger!" System 2 works more slowly and deliberatively. When balancing our checkbooks, for instance, it is System 2 that allows us to crunch the numbers and see how much (or little) is in the bank. The contents of representations tokened in System 1 operations are not available upon reflection: when you peek over the cliff's edge, you experience fear, but your *awareness* of the conscious cognitive states triggered by that awareness – for example, the thought that you are plummeting to your death and that you've lived your life poorly, or asking yourself if your life insurance premiums are fully paid up – is a result of System 2 activities. Peering over the cliff also illustrates how System 1 and System 2 interact with one another. When you look over the edge, System 1 tells you "Cliff … danger!" but what might get activated in System 2 are thoughts that you've forgotten to pay your insurance premiums. Patterns of activation in System 1 generate patterns of activation in System 2, but your awareness of System 2 activities does not necessarily entail awareness of System 1 activities. Perhaps you can recognize after the fact that looking over the edge of the cliff led you to think about your insurance, but it's not guaranteed.

In some cases, System 2 struggles upstream against the flow of System 1 outputs. For instance, your perception of chicken noodle soup in a bedpan or of feces-shaped fudge will quickly and automatically activate your feelings of aversion and disgust. But you can sometimes overcome these initial impulses. You can tell yourself that eating soup out of the bedpan is perfectly safe since the bedpan was just removed from its factory packaging, or that the chocolate fudge is only shaped like feces (Gendler, 2008, pp. 639, 640). In other cases, System 2 works to justify the results of System 1. That is, System 1 churns out a response and System 2 tries to justify the response. This is perhaps the best explanation of prestige bias – a tendency to favor research from well-known individuals and institutions.[13] In Peters and Ceci (1982), the experimenters resubmitted previously published papers under fictitious names and fictitious, unprestigious institutional affiliations (e.g. Tri-County Institute for Human Potential). The papers were chosen from journals known not to practice anonymous review and then were resubmitted to those very same journals. Peters and Ceci reported that over 90% of the resubmitted papers were rejected, with reviewers' reports often citing massive methodological error as the reason for rejection.[14] Of course, reviewers are supposed to be experts in their fields and would not reject a paper without submitting acceptable reasons – acceptable both to themselves and other experts, presumably. But it would appear that the prestige of the fictitious authors and institutions attached to already published papers played a powerful role in reviewer decisions.

One reasonable explanation for the prestige bias is that System 1 churns out a representation that System 2 then tries to justify in deliberation. System 1 triggers a representation like "Tri-County Institute for Human Potential … substandard!" as an automatic reaction to the unprestigious-sounding institute name and System 2 generates reasons to back up the opinion in System 1. This interpretation is also suggested by similar findings: anonymized grading leads to higher marks for women (Bradley, 1984), CVs with male names are typically judged to be of higher quality than CVs with female names (Steinpreis, Anders, & Ritzke, 1999), and women are less likely to be judged as original in their philosophical research (Valian, 1998).

What this means is that the biased representations churned out by System 1 are taken up and used in System 2. And there is further evidence that System 2 tries to justify the outputs of System 1. Often, subjects confabulate an explanation for an automatic judgment (Nisbett & Wilson, 1977). For instance, a study by Uhlmann and Nosek (2012) found that insecure subjects confabulate an explanation by blaming their culture for their biased beliefs. The researchers asked subjects to rank on a scale from 1 (strong disagreement) to 7 (strong agreement) how much they agreed with statements like "I would laugh at jokes about minorities" and "I should laugh at jokes about minorities." Many subjects assigned higher numbers to the "would"-statements than the "should"-statements, strongly suggesting that they tokened racist thoughts more often than they thought they should. Next, the experimenters asked subjects to write *either* about a time when they failed to live up to one of their most important personal values (this is the "threat" condition) *or* about a time when they succeeded in living up to one of their most important personal values (this is the "affirmation" condition). Subjects in the *threat* condition were more likely to attribute the cause of their racist thoughts to their culture; subjects in the *affirmation* condition were more likely to attribute the cause of their racist thoughts to themselves. The upshot is that when subjects feel that their self-worth is threatened – as might plausibly happen when egalitarian-minded agents recognize their unconscious bias – they will confabulate an explanation to blame their biases on the influence of cultural forces.[15] But attributing racial bias to culture is the sort of rational, deliberative process we expect from System 2.

It isn't only System 1 that impinges on System 2 functioning: System 2 impinges on System 1 in at least some ways. For example, implementation intentions – saying "When X, Y!" – affects performance on System 1 tests of bias. So saying to oneself "If I see a Black face, I will think 'good'!" reduces the degree of bias one exhibits on tests of implicit attitudes (Stewart & Payne, 2008). But the decision to utter an implementation intention is clearly a System 2 process. Thus, System 2 processes can affect System 1 processes, though not in the same way that System 1 processes affect System 2 processes.[16]

To sum up: we have seen that System 1 activity affects System 2 processes and that System 2 activity affects System 1 processes. A consequence of System 2 impinging on System 1 is that epistemic agents are able to indirectly regulate the effects of System 1 processes. This will prove important below when we argue that Gendler's epistemic pessimism can be avoided by recognizing the implications of the fact that System 1 representations can fail to be activated.

3. Racism, fast and slow

We just distinguished between automatic System 1 and deliberative System 2 processes, and we saw that System 1 outputs affect System 2 processes and that System 2 outputs affect System 1 processes. Now we will use these two points about Systems 1 and 2 to distinguish "fast" and "slow" racial bias.

The term "bias" can be ambiguous between a process and a product. On the one hand, the term can pick out a disposition. Someone's dispositions are biased when there is some epistemic weight pulling her judgments in some direction.[17] On the other hand, "bias" can pick out the manifestation of a disposition. An epistemic state is biased when it's the outcome of a biased process.[18] When we use "bias" here, we mean *biased dispositions* unless otherwise indicated. We're concerned with how epistemic processes exhibit bias, not necessarily with the state of being biased.[19]

Let's compare two classes of experiments. One involves fast processing and the other involves slow processing. What they both reveal is that implicitly biased judgments occur both when we are mindfully attending to the evidence and when we are making fast, split-second decisions. By highlighting fast and slow processing in biased judgments, we find support for our reconstruction of Gendler's argument. And, as we'll show, the reconstructed argument helps to reveal a silver lining that was hidden in Gendler's original presentation.

In implicit attitude tests (IATs), one common measurement of implicit racial bias, subjects are presented with pairings of terms. One pair effectively picks out some social group, such as "Black," "White," "old," "young," "male," or "female." The other is a positively or negatively valenced term, like "pleasant," "terrible," "joy," "agony," "glorious," or "evil." Subjects are faster to identify "Black" with negatively valenced terms than positively valenced ones. This suggests that representations of "Black" are more strongly connected to negatively valenced representations than positively valenced ones; so, the activation of Black-representations automatically activates terrible representations (Greenwald, McGhee and Schwartz, 1998).

In another type of IAT, subjects are shown either the face of a White male or a Black male and then asked to identify quickly an ambiguous object as either a gardening tool or a weapon. Subjects exposed to the White face tend to say it is a tool; subjects exposed to the Black face tend to say it is a weapon (Payne, 2001, 2006).

In yet another test, subjects are asked to play a first-person shooter game. They are instructed to shoot all and only characters brandishing weapons. Subjects shoot at images of Black men with non-weapon objects in their hands more often than at images of White men with non-weapon objects in their hands (Correll, Park, Judd, & Wittenbrink, 2002).

All of these experiments involve fast, automatic responses: subjects are required to respond to inputs as quickly as possible to reflect associations of representations in System 1. No effortful deliberation is required. Let us call these cases of *fast* racial bias.

Other experiments require subjects to engage in slower, more deliberate processing. For instance, some experiments show that subjects tend to evaluate résumés with stereotypically Black names as inferior relative to résumés with stereotypically White names (Bertrand & Mullainathan, 2004). Experimenters sent résumés to employers advertising in the "Help Wanted" sections of newspapers in Boston and Chicago. Some résumés had Black-sounding names, such as "Lakisha" and "Jamal," while other résumés had White-sounding names, such as "Emily" and "Greg." The résumés with White-sounding names received 50% more callbacks than résumés with Black-sounding names. Put another way, Whites have to send out 10 résumés for one callback, whereas Blacks have to send out 15 résumés for one callback. Additionally, highly qualified White-name résumés received 30% more callbacks than poorly qualified White-name résumés; however, highly qualified Black-name résumés and poorly qualified Black-name résumés were largely lumped together. The researchers also note the employers in these experiments all self-identified as Equal Opportunity Employers.[20]

A version of this experiment was run recently "in the wild" by a highly educated Black woman from New Jersey. Yolanda Spivey had worked in the insurance industry for 10 years prior to being laid off, and when she applied for hundreds of jobs within the first few months of unemployment, she received no responses. Then Spivey began submitting application materials under the name "Bianca White." Spivey started a new profile on the employment website Monster.com, created a new email address, and modified her outgoing message to say "Bianca White" rather than "Yolanda Spivey." Importantly, she kept all of the employment and education information the same for both Bianca White and Yolanda Spivey. After one week, Bianca White received nine phone calls and seven emails requesting interviews. Yolanda Spivey – submitting all the same materials only under a different name – had received zero phone calls and two emails. That is, in using a "White"-sounding name, Spivey received 14 more offers for interviews that week than she did using her "Black"-sounding name.[21]

In cases like these, subjects evaluate a candidate on the basis of some criteria. The subjects are not under the sorts of time pressures that we noted in cases of fast racial bias, where subjects were required to react quickly to inputs. Instead, they have an opportunity to think and reflect on their judgments, reflecting System 2 activities. Thus, we will call these cases of *slow* racial bias, where subjects' systematic biases are the result of deliberation.

Fast racial bias is a type of System 1 activity. These automatic and fast effects short-circuit conscious awareness. The results of System 1 processes are automatically generated representations that are typically insulated from conscious control. So when primed with a Black face and exposed to an ambiguous object, agents see it as a weapon. This perception is not within their control: they cannot opt to see it as something else in the heat of the moment. But slow racial bias is a type of System 2 activity. These slow and effortful effects are routed through conscious awareness. The results of System 2 processes are representations produced by conscious, deliberative processes. Subjects weigh the evidence they have in front of them and decide which candidate to hire. Fast and slow racial bias are distinguished with respect to the systems involved. Fast bias involves *only* System 1; slow bias involves *both* Systems 1 and 2.

What is the difference between the slow cases of bias and cases of overt racism? It is a matter of self-knowledge at least in part (cf. Kelly & Roedder, 2008). When individuals are overtly biased, they are transparently racist – they consciously entertain racist thoughts and endorse judgments motivated by racial discrimination. But in slow, System 2 discrimination, individuals unknowingly make biased decisions in virtue of System 1 processes affecting System 2 functioning. Slow bias is difficult to identify because the agent's biased behavior is apparent only when he or she reflects on patterns of behavior: introspection does not reveal the contents of System 1 representations affecting the System 2 processes. The biased agent offers reasons she endorses for her judgment, and unlike overt racism, there is no consciously entertained racist intention. So the only evidence available for System 2 bias are ongoing patterns of discriminatory behavior – the White female professor who almost exclusively calls on White female students for answers in class discussions, academic reviewers who prefer abstracts attached to stereotypically male rather than female names (Knobloch-Westerwick et al., 2013), or the auto mechanic who charges women more than men (Busse et al., 2015).

The experiments we've considered so far identify patterns of bias that are unrecognized by the epistemic agent. As we have noted, there is a crucial difference between experiments revealing fast racial bias and experiments revealing slow racial bias. In fast cases, subjects who complete IATs are told to go as quickly as they can; these behaviors are presumed to reflect connections forged in the subconscious System 1.[22] But slow cases find subjects taking their sweet time to choose the best candidate for the job. When a decision is made in slow cases, epistemic agents presumably endorse some supporting reasons but fail to recognize that their reasons reflect systemic bias.

But how is System 1 connected to fast bias, and how is System 2 connected to slow bias? The answer has to do with the nature of System 1 and 2 representations. On the one hand, System 1 representations are closely tied to affective states and behavioral routines. Gendler helpfully underlines the point with her discussion of a cognitive state she calls "alief." She writes that "a paradigmatic *alief* is a mental state with associatively linked content that is representational, affective, and behavioral, that is activated … by features of the subject's internal or ambient environment" (2008, p. 642).[23] System 1 representations have affective content and behavioral routines built right in. And since System 1 representations are tokened automatically, their effects fall outside the agent's direct control. As a result, the contents of System 1 representations of fast racial bias are insulated from agents' considerations of the epistemic weight given to the representations. On the other hand, System 2 representations are *not* closely tied to affective states or behavioral routines. (How often do people feel disgust or fear when reviewing résumés?) Instead, in cases of slow racial bias, agents effortfully assign epistemic weight to evidence. Even when agents' weightings are skewed by System 1 representations, the weightings are not automatically enacted in System 2 representational contents. Since System 2 representational contents are not closely tied to behavioral routines or affective states, these contents require assent from agents in order to play an epistemic role. But when agents make explicit judgments, putatively on the basis of their evidence, the influence of bias won't normally be apparent to them: since biased System 1 representations feed into the biased System 2 representations, agents remain oblivious.

4. The localist view: rejecting Gendler's epistemic pessimism

Let us return to Gendler's argument. Gendler's epistemic pessimism is the idea that living in a society structured by race entails epistemic costs, whether those costs are base rate neglect, constant vigilance to manage chains of associations, or implicitly endorsing claims that are explicitly rejected.

In our discussion of Gendler's reconstructed argument, we mentioned that there are important differences between System 1 costs and effects and System 2 costs and effects, and that the various pernicious effects of living in a racially structured society do not map neatly onto a dilemma of "to encode or not to encode." We also noted that epistemic agents are capable of encoding associations while avoiding the cascade of effects associated with that chain of associations. Studies in stereotype threat suggest as much: if the stereotype is not activated, then the stereotype-specific effects do not manifest themselves. We also find a related lesson in the psychological literature on taste aversion and learning. For example, Bernstein and Webster (1980) studied learned taste aversions. They offered two distinctively flavored ice creams – Maple Nut and Hawaiian Delight – to a control group and to a group undergoing chemotherapy. The cancer patients were given the ice cream just prior to their scheduled chemotherapy session. Each group was then approached a second time with the same ice cream options. This second time there was a twist: subjects were asked to taste both ice creams and pick whichever they preferred. For the control group, there was no clear correlation between the first round of tasting and the second. For the subjects undergoing chemotherapy, subjects tended to prefer whichever flavor they were *not* exposed to prior to treatment. Bernstein and Webster conclude that the subjects had developed a learned taste aversion to whichever flavor of ice cream they had tasted prior to chemotherapy. Subjects encoded an association between the distinctively flavored ice cream and feelings of nausea. If representations encoding the taste of the ice cream are not activated, then those feelings of nausea brought on by the taste of the ice cream are not experienced.

What we will call the *localist view* takes it cues from these features of our experience and the experimental literature. According to the localist view, the negative effects of living in a racially structured society can be reduced through the design of an agent's local environment. By changing an agent's local environment, the epistemic costs of living in a society structured by race can be undercut. Our localist view does *not* hold that epistemic agents will fail to encode representations of racial norms that are salient in the culture. Instead, the localist view proposes that those internalized representations of racial norms need not be activated. A bit more carefully: if implicitly biased behavior requires activation of wide networks of System 1 representations, then sufficiently large portions of the network need not be activated and thus will fail to generate implicitly biased behavior. Let us explain.

The agent's local environment includes the social and cultural representations she is subject to. That includes everything from TV shows to music, YouTube videos to billboards, and college textbooks to Sears catalogs. The local environment also includes other agents: friends, family, coworkers, taxi drivers, postal workers, and so on. Social psychology abounds with studies in which the local environment influences decision-making without the subject's awareness. One study by Harris, Bargh, and Brownell (2009), for example, found that exposure to food commercials increased snacking on available foods. Children watching television featuring food commercials, in particular, ate up to 45% more food than children watching television without any food commercials. Other agents are a particularly important part of an agent's environment. A wide range of studies shows that people tend to adopt attitudes of others in their social network (Christakis & Fowler, 2008; Cacioppo, Fowler, & Christakis, 2009).

The localist view emerges from careful consideration of the nature of System 1 and System 2 representations. System 1 and System 2 consist in networks of representations that are activated either through exposure to environmental stimuli or through activation of a connected representation. System 2 operates as it does, in part, because of System 1 processes. So if System 1 cranks out representations that promote bias (e.g. "Black dude … danger!") and those representations are taken up into System 2 activities, then System 2 activities will reflect the biased associations we find in System 1. *But if those biased associations are not activated in the first place*, then we would expect that System 2 will not reflect bias as strongly as it would have otherwise.[24] To repeat: failures to activate bias-promoting System 1 representations undercut bias-promoting System 2 processes.[25]

There is ample experimental evidence supporting this idea. We will briefly consider three cases. For a start, agents can manipulate their own performance on various measures of implicit bias by using implementation intentions – utterances by agents of the form "if X, Y!" In some studies, subjects were asked to pair images of Blacks with positive words (e.g. "enjoyable") or negative words (e.g. "terrible"). Subjects who said to themselves, "Whenever I see a Black face, I will think 'good'!" sharply curtailed the effects of their own implicit biases on their performance. In a second kind of case, perceiving the image of a counter-stereotypical exemplar – Martin Luther King Jr. or Sojourner Truth, for instance – has a similar effect of reducing the likelihood of exhibiting implicitly biased behavior as evidenced by scores on IATs.[26] Implementation intentions and exposure to counter-stereotypical exemplars help prevent activation of bias-promoting System 1 representations, thereby reducing the likelihood of agents unknowingly manifesting slow racial bias.

Finally, Gaither and Sommers (2013) show that Whites with a non-White roommate are not as averse to encounters involving non-Whites, relative to their White peers with White roommates. Whites with non-White roommates had a more racially diverse group of friends, exhibited reduced anxiety during inter-race interactions, and were more pleasant during an encounter with a Black individual they hadn't met before. The evidence suggests that common patterns of behavior exhibited by Whites can be undercut if they live with a non-White roommate. For our purposes, the crucial point is this: Whites living in a racially structured society do not *unavoidably* judge and act in implicitly biased ways. Instead, Whites' cognitive and behavioral dispositions around non-Whites can be shaped by the environment. (Admittedly, it's more difficult for many White people to set themselves up with a roommate than to look often at a picture of Martin Luther King, Jr.)

It isn't merely the *encoding* of racial information that generates biased judgments and behaviors. What matters is the encoding *and activation* of the relevant representations that generate biased judgments and behaviors. Encoding does not entail activation, as we have noted: racial information may be encoded, but it need not always be accessed. Encoded representations may be prevented from being activated in a number of ways. Implementation intentions and perception of counter-stereotypical exemplars can prevent the activation of the relevant patterns of representations needed for performing implicitly biased behaviors. So, avoiding the effects of living in a racially structured society involves following strategies uncovered by social psychologists for reducing the effects of implicit bias.

Our localist view bears some resemblance to views set forth by Sarkissian (2010) and Madva (2016). According to Sarkissian, one lesson to be learned from situationism in social psychology is that some environments promote praiseworthy actions while others fail to do so. He mentions the "seek/avoid" strategy: agents (morally) ought to seek out virtue-inducing environments and avoid virtue-debilitating environments. Our localist view differs from the seek/avoid strategy by emphasizing the *constructability* of local environments. It isn't just sometimes within our power to seek or avoid some environments. We can design new ones. Madva (2016) aims to address Gendler's dilemma by focusing on the role that implementation intentions can play in reducing implicitly biased judgments. Though we differ with Madva in how we understand Gendler's argumentative strategy, we concur with Madva in emphasizing the importance of representation activation (though he frames the matter in terms of *accessibility* of belief or knowledge). Madva's suggestion is that implementation intentions are a way that one can inhibit accessing biasing representations. Our view places greater emphasis on the construction of local environments in order to avoid activation of biased representations. As we see things, exposure to counter-stereotypical exemplars, increased interactions with minorities, and selective screening of the contents of one's environments are more permanent fixtures of bias-reducing environments than implementation intentions.[27] Here's another way to put our point: while Madva considers agent-centered solutions, we highlight agent- and environment-centered solutions.[28]

Now we turn to an objection to our localist view. Won't changing one's own local epistemic environment induce cognitive depletion? Change is hard. It takes effort to recognize how one's local environment promotes discrimination. It is perhaps even harder to do anything about it. Consequently, there are still unavoidable epistemic costs to be paid for living in a society structured by race, just as Gendler has argued.

In reply, we concede that it takes energy to make changes to one's epistemic environment, and this may lead to cognitive depletion of one's executive functioning. But this does not imply there are unavoidable epistemic costs for agents to pay. Here are two reasons.

First, executive cognitive functioning is like a muscle. The more it is exercised, the stronger it gets (Muraven & Baumeister, 2000). And just as it is unreasonable to expect a novice jogger to run an ultramarathon the first time out, it is unreasonable to expect biased agents to eliminate all sources of bias in the local epistemic environment in one shot. But just as novice joggers can train for the marathon, so too can egalitarian-minded epistemic agents plan to minimize the sources of bias in their local epistemic environments.[29] Thus, exhaustion in the short term, no matter whether it is jogging or executive functioning, brings greater endurance in the long-term.[30] Paying the epistemic costs here and now reduces costs that might be incurred later on. It is also worth observing that this strategy employs a "ratchet effect" (Tomasello, 1999): the effect of small, incremental changes adds up to significant differences, increasing the fund of cultural knowledge and thereby minimizing "slippage" back into earlier states. Changing one's local environment has initial start-up costs, true enough, but these are short-term costs, and the resulting changes to the environment promote further strengthening of one's executive function by having the environment bear some of the cognitive costs (Clark & Chalmers, 1998). A second reason why cognitive depletion upon reorganizing one's environment is not worrisome is that the cost need not be *personal*. The epistemic agent may be fortunate enough to live and work together with people who cover the cognitive costs to design and build the right sorts of environments. This means the agent enjoys the benefits of bias-reducing environments without paying any cognitive costs herself.

5. Conclusion

Our localist view holds that System 1 representations encode discriminatory information in the environment, but those representations affect agents' cognitive processes only when the representations are activated. Consequently, people can construct their local environment so that bias-inducing System 1 representations are not activated. Though the discriminatory information is encoded, it remains inactive. Thus, we can avoid paying the epistemic costs that Gendler has identified.

We conclude that there is good reason to reject Gendler's epistemic pessimism. It is not true that we must, one way or another, face the epistemic costs of implicit bias, so long as we can build local epistemic environments that do not activate those chains of bias-inducing representations.

Karl Marx once said that "philosophers have hitherto only *interpreted* the world in various ways; the point is to change it." Work on contemporary epistemology may not often be guided by philosophical ideals like that. And yet recent discussions concerning the effects of racial (and gender and social class) categories prompt us to do more than merely interpret the world. Following Gendler, our discussion has suggested how the goal of philosophically-guided cultural change might be reasonably pursued. We have suggested one way for philosophers to take seriously their commitments to egalitarian values. By understanding the nature of racial categories in our environments and their effects on minds, we can begin to see how to remake our epistemic worlds and ourselves for the better.

Notes

1. Over the last several years, a number of philosophers have sought to understand the philosophical implications of implicit bias research. For an overview of recent discussions, see Brownstein (2015). Issues concerning responsibility for biased judgments are taken up by Holroyd (2012, 2015), Saul (2012a), and Crouch (2012). In an epistemological vein, Puddifoot (2016) and Saul (2012b) argue that implicit bias challenges various theses concerning epistemic justification. A cluster of issues at the interface of epistemology and the psychology of implicit bias are explored by Sullivan-Bissett (2015), Holroyd and Sweetman (2016), and Mandelbaum (2015). Edited volumes by Brownstein and Saul (2016a,b) explore metaphysical, psychological, epistemological, moral, and political issues raised by implicit bias.
2. Another epistemic cost discussed by Gendler is cross-race facial deficit. This phenomenon is characterized by difficulties in distinguishing faces of people of different races: it is easier for Whites to distinguish individuals among White faces but difficult to distinguish individuals among Black faces. Mugg (2013) argues that this doesn't constitute an *epistemic* cost. We won't rehearse the details of Mugg's argument here, nor will we weigh in on whether cross-race facial deficit is a genuine epistemic cost. For our purposes, we can set aside discussion of cross-race facial deficit and focus on the other epistemic costs Gendler discusses.
3. A successful replication of this experiment by Gibson, Losee, and Vitiello (2014), which administered a questionnaire immediately after the math test, strongly suggests that subjects who were familiar with stereotypes about the math abilities of Asians and women were unaware that the priming affected them.
4. Shih and colleagues (1999) note that this is especially striking, because there is also a stereotype that Asians are good at math. In fact, when they primed Asian-American female students to reflect on languages spoken at home, their performance on a subsequent math test improved relative to those students who weren't asked to reflect on languages spoken at home.
5. Spencer, Steele, and Quinn (1999) suggest that the testing situation itself is enough to activate the stereotype and decrease performance.
6. Egan (2011) and Mugg (2013) both recognize D1 as a crucial step in Gendler's argument.
7. D2 leaves open the possibility that agents fail to encode information about inequality in the first place. That's possible but downright unlikely given the ubiquity of representations of inequality in our society.
8. Koehler (1996) likewise distinguishes between representations of base rates, but as a result of implicit vs. explicit learning. What we are calling "implicit" discounting of base rates maps onto Koehler's "direct experience" of base rates that leave a "trace" in the representational system. Given enough traces, the information becomes cognitively available. This "contrasts with the explicit learning of a single summary

statistic that does not produce multiple traces and is associated with less accurate judgments" (1996, p. 7). Gendler acknowledges these complexities (p. 37, Note 6), but we think that they are more relevant to the discussion than she suggests.

9. Could we say that the cost of (1), base rate neglect is mapped to the cost of (4), implicit irrationality and the cost of (2), association regulation is mapped to (3), explicit irrationality? No. Costs (1) and (3) are results of encoding racial categories while (2) and (4) are results of failing to encode. It's not possible that equivalent costs are results of both encoding and failing to encode racial categories.
10. Here's an objection to premise 6: there is evidence that merely encoding the biasing information has epistemic costs. For example, Hahn, Judd, Hirsh, and Blair (2014) show that subjects are surprisingly accurate at predicting outcomes of tests of their implicit attitudes. Consequently, merely encoding (but not activating) implicit attitudes carries an epistemic cost, at least for some agents: trying to avoid activation of biasing representations. The objection runs together two ways in which biasing information can generate epistemic costs: direct and indirect. Biasing information *directly* generates epistemic costs when activation of representations of that information causes an agent's judgments to be less reliable, causes an agent to become cognitively depleted, and so on. Biasing information *indirectly* generates epistemic costs when unactivated representations of that information cause an agent's judgments to be less reliable, cause an agent to become cognitively depleted, and so on. Take an analogy. Smith's car swerved because she turned the steering wheel hard; Smith is the direct cause of the car's movement, but Smith turned the wheel hard to avoid the pothole; the pothole is the indirect cause of the swerve. The study by Hahn and colleagues (2014) shows that encoded information about race can indirectly generate epistemic costs. An agent's awareness of her bias can cause her to take preventative measures against its activation. But in this case, the biasing information itself does not generate epistemic costs; instead, the awareness of the implicit bias directly generates the costs. Thus, awareness of implicit bias directly generating epistemic costs is a case distinct from biasing information indirectly generating epistemic costs. In what follows, we are concerned only with how biasing information directly generates epistemic costs. (Thanks to an anonymous reviewer for helpful discussion of the objection.).
11. For an overview of dual-process theories, see Evans (2008) and Frankish (2010). Kahneman (2011) is a book-length overview of his work on dual-process theories. Evans and Stanovich (2013) review major objections to dual-process theories. They argue that there is no generic dual-systems account: different accounts highlight different mechanisms, capacities, and properties. Consequently, many of the objections to dual-systems in general fail. We do not propose to argue that there is some general account. Our aim is only to identify a handful of properties that seem to hold of Systems 1 and 2. The review works mentioned above fit with our characterization of Systems 1 and 2. As far as we know, there is yet to be a case of a conscious System 1 process or an automatic System 2 process. We thank an anonymous reviewer for pressing us on this point.
12. Two common architectures for how the two systems work are parallel-competitive processing and default-interventionist processing. The former suggests that Systems 1 and 2 work in parallel and that the outputs of each jostle for position within the cognitive system (e.g. Epstein, 1994). The latter suggests that judgments are usually the product of System 1 unless System 2 overrides the outcome of System 1 (e.g. Kahneman, 2011).
13. See Peters and Ceci (1982) and Saul (2012a).

14. Saul (2012a, Note 7) notes that this reason for rejection is atypical in psychology journals (also see Lee & Schunn, 2010). This suggests, perhaps, that unprestigious institutional affiliation plays some role in the dismissive attitudes of the reviewers.
15. Andreychik and Gill (2012) report that some agents will justify their own biased attitudes by a kind of empathy: in explaining their own biases, agents will appeal to the oppression of groups who are discriminated against. It's not clear, though, whether agents think of such "external explanations" as merely *explaining* their implicit bias or *normatively justifying* their bias. Andreychik and Gill suggest that explaining is most likely since at least some of the agents appealing to external explanations are motivated by compassion and empathy for targets of discrimination. On the other hand, subjects in Uhlmann and Nosek's threat condition may be motivated to engage in confabulation as a result of cognitive dissonance. How so? In the threat condition, subjects are asked to bring to mind a time when they failed. A reasonable inference is that subjects experience cognitive dissonance between thinking well of themselves and thinking poorly of themselves. Resolving the dissonance in this case involves attributing their failures to be egalitarian-minded to cultural influences. An anonymous reviewer made the interesting suggestion that these subjects' confabulated judgments are accurate; however, the accuracy of their judgments is irrelevant to their being confabulations. (Thanks to Keith Payne for the pointers to the literature and to an anonymous reviewer for the objection.).
16. A relevant question for further philosophical and psychological research is: How does the social environment impinge on System 2 processes? Consider this case of biased social representations and cognitive functioning: Blacks and Latinos are disproportionately represented as lawbreakers in television news (Bjornstrom, Kaufman, Peterson, & Slater, 2010). Plausibly, this contributes to forging connections in System 1 between representations of Blacks and Latinos and feelings of fear and danger. Now consider how these social messages affect System 2 processes. One explanation is that social messages about Blacks and Latinos indirectly affect System 2 via System 1: socially shaped System 1 processes affect System 2 processes. One consequence is that System 2 is untouched directly by social messages but only receives its information from System 1. Or living in a racially structured society might affect System 2 directly: a person might believe, as a consequence of seeing disproportionately more Blacks and Latinos as lawbreakers on television news, that a Black or Latino man in her neighborhood is more likely to commit a crime than a White man. How these elements are organized isn't as important for now as is highlighting that (1) living in a racially structured society has System 1 and System 2 effects, (2) System 1 and System 2 effects are distinct, and (3) System 1 effects impinge on System 2 processes and vice versa.
17. For example, Mandelbaum (2015) says that implicit attitudes are the result of unconscious beliefs – "honest-to-goodness propositionally structured mental representations that we bear the belief relation to" (p. 635). These unconscious beliefs eventuate in biased behaviors (cf. Mandelbaum, 2014). For related views, see Levy (2015), Machery (2016), and Schwitzgebel (2013).
18. Madva and Brownstein (in press; see also Brownstein & Madva, 2012) endorse this kind of view. They describe implicit states as "mutually co-activating semantic-affective-behavioral 'clusters' or 'bundles.'" Their position is similar to Gendler's, who describes implicit states as having affective, representational, and behavioral components.
19. Thanks to an anonymous reviewer for helpful comments on this matter.

20. Objection: aren't CV cases really instances of fast bias? Employers are sifting through applications and reaching decisions based on, for example, applicants' names. In picking "Emily" over "Lakisha," the employers rely on System 1, not System 2. We expect that employers in such cases offer reasons to rationalize their decisions or at least are disposed to do so. Such rationalizing makes typical résumé cases instances of slow racial bias. But even if we aren't correct about that, it does not follow that there are no cases of slow racial bias. Consider the growing literature on racial bias in jury deliberations. Given the length of time that jurors take to deliberate, it's reasonable to suppose that if there is bias in jurors' deliberations, it is slow bias. Sommers and Ellsworth (2000), for example, identify a range of cases in which Whites are liable to exhibit anti-Black bias in a courtroom setting, including interracial trials (e.g. a Black defendant and a White victim) where race is *not* a salient factor. In such cases, Whites are convicted at a rate of under 70% while Blacks are convicted at a rate of 85%. By contrast, in cases where race is a salient factor, the conviction rates for Whites and Blacks are both around 75%. (Thanks to an anonymous reviewer for suggesting we discuss this objection.).
21. http://www.techyville.com/2012/11/news/unemployed-black-woman-pretends-to-be-white-job-offers-suddenly-skyrocket/# Spivey doesn't report whether the emails and phone calls were from different employers. But even if every email were duplicated as a phone call – that is, if Bianca White received only nine requests for interviews – that still means that switching from "Yolanda Spivey" to "Bianca White" resulted in seven more interviews during the week she ran the experiment.
22. In fact, when subjects take their time on IATs, evidence of implicit bias goes down sharply (Fiedler & Bluemke, 2005; Cvencek et al., 2010).
23. Egan (2011) argues that we need not appeal to aliefs to account for the effects of System 1 representations.
24. Our position leaves open that System 2 processes can exhibit bias without activation of bias-promoting System 1 representations. It is possible that System 2 can harbor biased representations without input from System 1 – that's a fair description of the overt racist. But Gendler's discussion focuses on implicitly biased agents who are committed to egalitarian ideals. After all, what makes implicitly biased judgments so alarming is that they are made by people who are explicitly committed to egalitarian ideals. Consequently, we can safely assume that System 2 processes are not biased.
25. But isn't failing to activate relevant information a form of base rate neglect? Not always. In an example from the Central Intelligence Agency's *Psychology of Intelligence Analysis*, noted by Gendler (2011), subjects tended to neglect the ratio of Vietnamese to Cambodian jet fighters. Subjects in this case (as well as the ones described in Tetlock et al., 2000) failed to consciously appreciate information. But there are also cases in which subjects fail to consciously appreciate information but don't commit base rate neglect. Jones is baking a cake for a party and doesn't recall that one partygoer is lactose-intolerant. Even if Jones could have brought that information to mind, he is not thereby committing base rate neglect. Failing to consciously appreciate information is a necessary but not a *sufficient* condition for base rate neglect. We are suggesting that subjects will fail to consciously appreciate information when it is not activated, and thus remain unbiased, but that is not sufficient for counting as base rate neglect. (Thanks to an anonymous reviewer for discussion.).
26. For a survey of such data, see Lai et al. (2013).
27. Amodio nicely captures our concern: "implicit racial biases are particularly difficult to change in a cultural milieu that constantly reinforces racial prejudices and stereotypes" (2014, p. 679). Our suggestion is to change one's local sociocultural

milieu. We might add that Madva's conclusion coincides with Amodio's for mitigating the effects of implicit biases: focus on individual control-based interventions.

28. Note that our conclusions are similar to ones discussed in Haslanger (2015) and Fricker (2010). Haslanger argues that explanations of injustice invoking implicit biases are incomplete without appeal to social structures. We agree. But our position focuses on avoiding epistemic costs raised by implicit biases, not on explanations of injustice due to implicit biases.
29. Mugg (2013) offers a similar response to Gendler's worries about executive depletion.
30. No pain, no gain, as Jane Fonda once proposed.

Acknowledgments

For comments and conversations, we are grateful to Joseph Vukov, Shane Wilkins, Carlo DaVia, Peter Seipel, and anonymous referees. The first author would like to thank the Leverhulme Trust for funding the development of the Implicit Bias and Philosophy Research Network; the first author's discussions in the Network's workshops contributed to this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the development of the Implicit Bias and Philosophy Research Network [grant number IN-058].

References

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews Neuroscience, 15*, 670–682.

Andreychik, M. R., & Gill, M. J. (2012). Do negative implicit associations indicate negative attitudes? Social explanations moderate whether ostensible "negative" associations are prejudice-based or empathy-based. *Journal of Experimental Social Psychology, 48*, 1082–1093.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*, 991–1013.

Bernstein, I. L., & Webster, M. M. (1980). Learned taste aversions in humans. *Physiology & Behavior, 25*, 363–366.

Bjornstrom, E. E. S., Kaufman, R. L., Peterson, R. D., & Slater, M. D. (2010). Race and ethnic representations of lawbreakers and victims in crime news: A national study of television coverage. *Social Problems, 57*, 269–293.

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

Brownstein, M. (2015). Implicit bias. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Spring 2015 ed.). Retrieved from http://plato.stanford.edu/archives/spr2015/entries/implicit-bias/

Brownstein, M., & Madva, A. (2012). The normativity of automaticity. *Mind & Language, 27*, 410–434.

Brownstein, M., & Saul, J. (Eds.). (2016a). *Implicit bias & philosophy: Volume I, metaphysics and epistemology*. Oxford: Oxford University Press.

Brownstein, M., & Saul, J. (Eds.). (2016b). *Implicit bias and philosophy: Volume 2, moral responsibility, structural injustice, and ethics*. Oxford: Oxford University Press.

Busse, M., Israeli, A., & Zettlemeyer, F. (2015). *Repairing the damage: The effect of price expectations on auto repair quotes*. Working Paper #0126. The Center for the Study of Industrial Organization at Northwestern University. Retrieved April 1, 2015, from http://www.kellogg.northwestern.edu/faculty/directory/busse_meghan.aspx#research

Cacioppo, J. T., Fowler, J. H., & Christakis, N. A. (2009). Alone in the crowd: The structure and spread of loneliness in a large social network. *Journal of Personality and Social Psychology, 97*, 977–991.

Christakis, N. A., & Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *New England Journal of Medicine, 358*, 2249–2258.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 7–19.

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

Crouch, M. (2012). Implicit bias and gender (and other sorts of) diversity in philosophy and the academy in the context of the corporatized university. *Journal of Social Philosophy, 43*, 212–226.

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., & Snowden, R. J. (2010). Faking of the implicit association test is statistically detectable and partly correctable. *Basic and Applied Social Psychology, 32*, 302–314.

Egan, A. (2011). Comments on Gendler's, "The epistemic costs of implicit bias". *Philosophical Studies, 156*, 65–79.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist, 49*, 709.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspectives on Psychological Science, 8*, 223–241.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology, 27*, 307–316.

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass, 5*, 914–926.

Fricker, M. (2010). Replies to Alcoff, Goldberg, and Hookway on epistemic injustice. *Episteme, 7*, 164–178.

Gaither, S. E., & Sommers, S. R. (2013). Living with an other-race roommate shapes whites' behavior in subsequent diverse settings. *Journal of Experimental Social Psychology, 49*, 272–276.

Gendler, T. S. (2008). Alief and belief. *Journal of Philosophy, 105*, 633–663.

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

Gibson, C. E., Losee, J., & Vitiello, C. (2014). A replication attempt of stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999). *Social Psychology, 45*, 194–198.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*, 1369–1392.

Harris, J. L., Bargh, J. A., & Brownell, K. D. (2009). Priming effects of television food advertising on eating behavior. *Health Psychology, 28*, 404–413.

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

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy, 43*, 274–306.

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 M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy* (Vol. 1, pp. 80–103). Oxford: Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Macmillan.

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

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

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*(1), 1–53.

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass, 7*, 315–330.

Lee, C. J., & Schunn, C. D. (2010). Philosophy journal practices and opportunities for bias. *American Philosophical Association Newsletter on Feminism and Philosophy, 10*, 5–10.

Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs, 49*, 800–823.

Machery, E. (2016). De-Freuding implicit attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume I: Metaphysics and epistemology* (pp. 104–129). New York, NY: Oxford University Press.

Madva, A. (2016). Virtue, social knowledge, and implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, Vol. 1: Metaphysics and epistemology* (pp. 191–215), New York, NY: Oxford University Press.

Madva, A., & Brownstein, M. (in press). Stereotypes, prejudice, and the taxonomy of the implicit social mind. *Nous.*

Mandelbaum, E. (2014). Thinking is believing. *Inquiry, 57*, 55–96.

Mandelbaum, E. (2015). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs, 50*, 629–658.

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

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin, 126*, 247–259.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181–192.

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

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences, 5*, 187–195.

Puddifoot, K. (2016). Accessibilism and the challenge from implicit bias. *Pacific Philosophical Quarterly, 97*, 421–434.

Sarkissian, H. (2010). Minor tweaks, major payoffs: The problems and promise of situationism in moral philosophy. *Philosophers' Imprint, 10*(9), 1–15.

Saul, J. (2012a). Ranking exercises in philosophy and implicit bias. *Journal of Social Philosophy, 43*, 256–273.

Saul, J. (2012b). Skepticism and implicit bias. *Disputatio, 5*, 243–263.

Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelmann (Ed.), *New essays on belief: Constitution, content, and structure* (pp. 75–99). Houndmills: Palgrave MacMillan.

Shih, M., Pittinsky, T., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10*, 80–83.

Sommers, S. R., & Ellsworth, P. C. (2000). Race in the courtroom: Perceptions of guilt and dispositional attributions. *Personality and Social Psychology Bulletin, 26*, 1367–1379.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4–28.

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles, 41*, 509–528.

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

Sullivan-Bissett, E. (2015). Implicit bias, confabulation, and epistemic innocence. *Consciousness and Cognition, 33*, 548–560.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology, 78*, 853–870.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Uhlmann, E. L., & Nosek, B. A. (2012). My culture made me do it. *Social Psychology, 43*, 108–113.

Valian, V. (1998). *Why so slow? The advancement of women*. Cambridge: MIT Press.