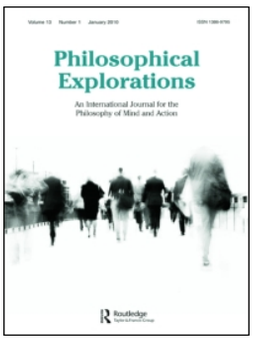

Routledge
Taylor & Francis Group

Philosophical Explorations

An International Journal for the Philosophy of Mind and Action

ISSN: 1386-9795 (Print) 1741-5918 (Online) Journal homepage: http://www.tandfonline.com/loi/rpex20

Dissolving the epistemic/ethical dilemma over implicit bias

Katherine Puddifoot

To cite this article: Katherine Puddifoot (2017) Dissolving the epistemic/ethical dilemma over implicit bias, Philosophical Explorations, 20:sup1, 73-93, DOI: 10.1080/13869795.2017.1287295

To link to this article: https://doi.org/10.1080/13869795.2017.1287295

Published online: 11 Apr 2017.

Philosophical Explorations, 2017
Vol. 20, No. S1, S73–S93, http://dx.doi.org/10.1080/13869795.2017.1287295
This paper is part of the online only special issue "False but Useful Beliefs" guest edited by Lisa Bortolotti and Ema Sullivan-Bissett.

Dissolving the epistemic/ethical dilemma over implicit bias

Katherine Puddifoot*

Philosophy, University of Birmingham, Birmingham, UK

(*Received 22 November 2016; final version received 5 December 2016*)

It has been argued that humans can face an ethical/epistemic dilemma over the automatic stereotyping involved in implicit bias: ethical demands require that we consistently treat people equally, as equally likely to possess certain traits, but if our aim is knowledge or understanding our responses should reflect social inequalities meaning that members of certain social groups are statistically more likely than others to possess particular features. I use psychological research to argue that often the best choice from the epistemic perspective is the same as the best choice from the ethical perspective: to avoid automatic stereotyping even when this involves failing to reflect social realities in our judgements. This argument has an important implication: it shows that it is not possible to successfully defend an act of automatic stereotyping simply on the basis that the stereotype reflects an aspect of social reality. An act of automatic stereotyping can be poor from an epistemic perspective even if the stereotype that is activated reflects reality.

Keywords: Implicit bias; stereotypes; stereotyping; epistemic innocence; epistemic costs and benefits; distorted judgement

1. Introduction

Imagine the following scenarios. Pete and Anita have both developed into adulthood in societies in which women but not men are provided with parental leave after the birth of a child. In these societies, because of the parental leave provision, women and not men tend to stay at home to look after the children just after birth. Out of necessity and often through trial and error, women tend to develop characteristics appropriately described as nurturing while male partners do not develop these characteristics because they do not occupy the position of primary caregiver.[1] However, there is nothing in the nature of women that means that they are nurturing and many men are nurturing. Pete is exposed to a media portrayal of family life in which men are predominantly cast in a provider role while women are cast as nurturing. Women are therefore portrayed as if they will be better at childrearing than their male counterparts. Anita, in contrast, has been brought up in a society in which men and women are portrayed in the media as equally nurturing, and equally good at childrearing. While Pete becomes disposed to automatically respond to women as if they are more nurturing than men, Anita becomes disposed to automatically respond as if men and women are equally nurturing.

*Email: k.puddifoot@bham.ac.uk

Now consider another example. Farah and Timothy have been brought up in a society in which women are severely underrepresented at senior levels in the sciences. In the societies in which they live there are deeply entrenched and systematic inequalities in science meaning that many women hit a glass ceiling before they reach senior levels. Both Farah and Timothy have consequently both become predisposed to automatically respond as if all senior scientists are male. When told that they will meet a senior scientist, they both automatically respond as if the scientist will be male; when told about the latest advanced feat of engineering they automatically picture a male engineer leading the team behind it, and so on. Both Farah and Timothy are on hiring panels, selecting candidates to interview for a junior role as a scientific researcher. The aim in both cases is to identify a future leader in their scientific field. Farah ensures that the application material she considers is anonymous, with the gender removed, so that she does not respond differently to the applications of men and women. She does not automatically associate the male candidates more strongly than the female candidates with advanced scientific expertise and leadership because she is unaware of the candidates' gender. In contrast, when Timothy views the application material he has access to information about the gender of the candidates. He consequently automatically associates the male candidates more strongly than the female candidates with advanced scientific expertise and leadership, reflecting the social reality that scientific experts are more likely to be male than female.

Now imagine that you can determine whether someone you care about is brought up in a situation like that of Pete or Anita, or makes a recruitment decision using anonymous or non-anonymous application material. You are choosing for that individual to either be disposed to automatically respond in a way that reflects the social realities, or to automatically respond in what shall be described in this paper as the *egalitarian way*. The automatic responses that reflect the social realities include responding as if any nurturing person is more likely to be a woman than a man and responding as if any future leader in science is more likely to be male than female. Meanwhile, the egalitarian responses include responding as if any nurturing person who is good at childrearing is equally likely to be a man or a woman; and responding to candidates of each gender as if they are equally likely to be future scientific leaders. It seems as if you face an ethical/epistemic dilemma (Kelly and Roedder 2008; Haslanger 2010; Egan 2011; Gendler 2011; Mugg 2013; for discussion see Brownstein 2015): you can choose either (i) the ethical choice, which is for the individual to respond in the egalitarian way; or (ii) the epistemically beneficial choice, that is, the choice that will increase the chance of the individual forming true beliefs and making accurate judgements. On this view, the ethical choice would make the individual less likely to achieve epistemic goals like obtaining true beliefs or fitting their beliefs to the evidence. It might seem, in other words, that to achieve goals such as obtaining true beliefs or fitting beliefs with the evidence, one should *reflect the social realities of inequalities* in one's automatic responses.

In this paper I challenge this intuitive position, arguing that failing to reflect realities of inequalities in our automatic responses can be the best thing from an epistemic perspective. To articulate this idea, I utilise the notion of *epistemic innocence* that has been introduced by Bortolotti (2015a, 2015b) and Sullivan-Bissett (2015) (see also Letheby 2016). A cognition is epistemically innocent if it has obvious epistemic costs but also brings significant epistemic benefits that could not have been gained in the absence of the costly cognition. I recognise that there can be epistemic costs when humans fail to reflect social inequalities such as the underrepresentation of women in science in their automatic responses. However, I argue that cognitions that fail to reflect realities in this way can also bring significant epistemic benefits that could not have been gained if the judgements reflected the social realities. Furthermore, the epistemic benefits of a cognition that fails to reflect social realities,

and the fact that these benefits cannot be obtained with a cognition that properly reflects social realities, means that the cognition is often the lesser of two epistemic evils. Consequently it is often the case that the best choice from an epistemic perspective is the same as the ethical choice: to fail to reflect social inequalities in one's cognition.

In order to establish the epistemic benefits that can follow from failing to reflect social realities in one's judgments, I show that there are substantial epistemic costs to the alternative: engaging in what social psychologists describe as stereotyping. In the psychological literature, the standard viewpoint is that stereotypes are associations between characteristics, attributes, and behaviours and certain social groups (see, e.g. Hilton and von Hippel 1996, 240; Beeghly 2015). According to this definition, when we stereotype we associate certain characteristics, attributes and behaviours more strongly with the social group about which we stereotype than other groups. There is an alternative, normatively loaded definition according to which stereotypes are inaccurate (see, e.g. Blum 2004), but this is not the definition that is utilised within psychology. This means that when psychologists discuss the impacts of stereotyping they are referring to the consequences of associating members of some social groups more strongly than others with certain traits, including cases in which the association reflects some aspect of social reality. The cognitions that the psychological literature on stereotyping describes therefore include those that are of interest in the current discussion, that is, cognitions that associate group members with traits that members of their group are statistically more likely than members of other groups to possess. For ease of discussion I adopt the descriptive rather than the normative definition of stereotypes advocated within social psychology, so henceforth the term *stereotype* will be used to capture all associations of members of certain groups more strongly than others with characteristics, attributes or behaviours. As my interest is social stereotypes, I use the term to refer to associations with members of social groups.

It should be noted that my claim is not that on all occasions when we fail to reflect social inequalities in our judgements the resulting cognitions are the best from an epistemic perspective. Nor is my claim that when we associate social groups with certain traits or characteristics we always or even often reflect social realities in doing so. There is good reason to think that many associations of this type inaccurately represent social realities (Puddifoot forthcoming). I argue for two claims that are more modest than this. First, that our cognitions can be epistemically innocent when we automatically respond in an egalitarian way, treating individuals from different social groups (e.g. women/men) as though they are equally likely to possess a trait (e.g. being a scientist, being nurturing), *even under those conditions* in which members of the relevant social group are statistically more likely to possess the trait. Second, that egalitarian cognitions of this sort are *often* epistemically innocent in this way.[2]

Although these claims are circumscribed, they are important for a number of reasons. First of all, they show that we should not overstate the case for the epistemic/ethical dilemma. In spite of appearances, in many situations we should not raise objections to the ethical option on the basis that it has epistemic costs, or vice versa, because the same option is likely to be the best from both the epistemic and ethical perspective. Second of all, and relatedly, the epistemic innocence claim provides a way to respond to one natural defence of automatic stereotyping. When people are criticised for stereotyping on the basis that they are being unethical, they might respond that they are just following the facts, responding in the way that those who want knowledge or understanding should respond, because the stereotypes that they are applying reflect social reality. The argument in this paper shows that it is possible to respond to this claim in more than one way. If one has the required evidence to hand, one can show that the stereotype does not truly reflect reality. But even if one does not have this evidence, based on the arguments

in this paper one can present reasons for thinking that, as a result of stereotyping, the agent is ultimately less likely to follow the facts because of the downstream epistemic costs of stereotyping which often follow regardless of the way that the stereotype reflects reality.

Third, the claims in this paper provide implications for how we should think about strategies to mitigate the effects of implicit bias. Implicit biases are "fast, automatic, and difficult to control processes that encode stereotypes and evaluative content, and influence how we think and behave" (Holroyd and Puddifoot, forthcoming). Implicit biases lead to the association of members of certain social groups with concepts or affective responses, operate unintentionally and can occur outside the agent's awareness.[3] If people automatically respond in the way that reflects social realities, they will be influenced by implicit biases that lead some social groups to be associated more strongly than others with certain, often desirable characteristics. Various strategies have been advocated to reduce the influence of implicit bias on thought and action, and, as proponents of the ethical/epistemic dilemma have pointed out, many of the strategies have the potential to prevent information about social realities from influencing our judgements. Some strategies that have been advocated involve preventing associations from being triggered. For example, the strategy outlined in the second opening example of ensuring that curriculum vitae are anonymous at the point that they are evaluated has been advocated as a way to reduce the impact of implicit bias (Steinpreis, Anders, and Ritze 1999; Saul 2013). If this strategy is implemented, information identifying the social group membership of a candidate, such as their gender, age, religion, ethnicity or details about any disabilities they have, is removed. When people make judgements of anonymous curriculum vitae their judgements are unlikely to be influenced by automatic associations with a social group because the evaluators will not know which social group the candidates belong to. In another example, the use of *implementation intentions* or "if-then" plans has been encouraged (e.g. Stewart and Payne 2008; Saul 2013; Madva 2016a). When someone forms an implementation intention they form a specific action plan such as *if I meet a woman in science, I will think scientific expert.* The plans are specified so that they lead to the triggering of specific concepts or affective responses (e.g. the concept *expert* and/or a positive affective response) in the presence of specific cues (e.g. a woman in science). Where a person would make an association that better reflected the social reality in the absence of the implementation intention, for example, an association between scientific expertise and men, the implementation intention will prevent associations that reflect something of social reality from being triggered and influencing a judgement. A third strategy to mitigate implicit bias involves changing the associations that people make with members of certain social groups. People are encouraged to consider counter-stereotypical individuals, that is, individuals who do not fit the stereotype of a social group, to change the way that they respond to members of the group (Blair, Ma, and Lenton 2001). The strategy of considering counter-stereotypical examples can be viewed as leading people to less accurately reflect social reality in their judgements where the initial association reflected some aspect of social reality. It might therefore be thought that each of these strategies used to counter implicit bias is problematic from an epistemic perspective whenever our initial automatic responses would reflect social reality because they lead people to make judgements that fail to reflect the social realities: either by controlling the triggering of the associations or changing the associations that are made. A final implication of the argument in this paper is that even where the strategies have the consequence of leading people to fail to reflect social reality they can nonetheless be viewed positively from an epistemic perspective.[4]

2. Epistemic innocence

The notion of epistemic innocence is central to the current discussion, so this section outlines in more detail what it is for cognition to be epistemically innocent. Bortolotti (2015a) introduces the notion of epistemic innocence through a comparison with the *justification-defence* in UK and US law. A justification-defence is used to establish the innocence of an individual by showing that their act did not constitute an offence under the circumstances in which it was performed. Justification-defences can apply where an action prevents serious harm. They apply where an action that would be viewed as a criminal offense under other circumstances is viewed as an emergency response. Bortolotti's claim is that a similar defence can be made of some epistemically costly cognitions. While they bring epistemic costs, they lead to the avoidance of other epistemic costs that could not otherwise have been avoided. Two conditions are met by any epistemically innocent cognition:

(A) *Epistemic benefit*: the cognition confers some significant epistemic benefit to an epistemic agent at a given time.
(B) *No Alternatives*: alternative cognitions that would confer the same epistemic benefit without the epistemic costs are not available to the agent at that time.

My claim is that cognitions that fail to reflect social inequalities are often epistemically innocent. While they bring epistemic costs, they lead to the avoidance of other significant epistemic costs. By leading to the avoidance of these costs, the cognitions confer significant epistemic benefits to the agent at the given time that could not have been obtained if the agent had a cognition that was non-faulty in the sense that it reflected the relevant social inequalities.

A subset of epistemically innocent cognitions are the *lesser of two epistemic evils*; while they are faulty because they bring epistemic costs, they lead to the avoidance of more substantial epistemic costs associated with cognitions that lack their faults (Bortolotti 2015a). In addition to showing that cognitions that fail to reflect social inequalities can be epistemically innocent, I aim to show that the same cognitions often belong to the subset of epistemically innocent cognitions that are the lesser of two epistemic evils.

I first show that there can be epistemic benefits to making automatic associations between certain social groups and their members and certain traits, and therefore epistemic costs to failing to make these associations (section 3). Then I show that these costs can be outweighed by the benefits of failing to make the associations, so that cognitions that fail to reflect social inequalities can be epistemically innocent. To achieve the latter goal, I show the following things:

(1) There are epistemic costs of reflecting social inequalities in our judgements (section 4).
(2) These costs can be avoided by failing to reflect social inequalities in our judgements (section 5).
(3) There are often no alternative ways (other than failing to reflect social inequalities) to gain the epistemic benefits of failing to reflect social inequalities in our judgements (section 6).
(4) The costs of failing to reflect social inequalities in our judgements are often outweighed by the benefits (section 7).

An important question is raised by any discussion of epistemic costs and benefits: what exactly makes something a cost or a benefit? In previous discussions of epistemic costs and

benefits, a consequentialist approach has been taken (Bortolotti 2015a; arguably also Gendler 2011) and a cognition is viewed as epistemically good if it leads to positive consequences from an epistemic perspective while something is viewed as epistemically bad if it leads to negative consequences. I too focus on consequences, arguing that there are positive consequences that follow from failing to reflect some social realities in our judgements: an agent becomes disposed to respond in ways that increase the chance that they will obtain true beliefs and become disposed to form beliefs that are more consistent with and better supported by the evidence available to them. It is worth noting, however, that the claim that a cognition is epistemically innocent is not equivalent to the claim that the believer is justified in believing the products of the cognition. It is consistent with the discussion in this paper, for example, that to be *justified* in making the choice not to reflect certain social realities in one's judgements, and in believing the product of the judgement, one might need to be aware of the epistemic innocence of cognitions that fail to reflect the realities.

3. Epistemic benefits to accurately reflecting reality

The aim of this paper is thus to establish that the epistemic benefits of failing to reflect social realities in our automatic judgements often outweigh the epistemic costs. My focus is therefore predominantly on the epistemic benefits of failing to reflect social inequalities in our automatic responses, but as I argue that failing to reflect social realities can be the lesser of two epistemic evils, it is worth briefly providing a case in support of the idea that it is epistemically costly. As a consequentialist approach is being taken to epistemic costs and benefits, to establish that there are epistemic costs, it will suffice to show that we are disposed, as a consequence of failing to reflect social realities, to respond poorly from an epistemic perspective under some conditions. What follows, then, is an example that shows one way that we can be disposed to respond in ways that will reduce the chance that we will make accurate judgements as a result of failing to reflect social realities in our judgements.

Imagine that you are asked whether a randomly selected female is more likely to be a scientist or a non-scientist. You harbour an implicit bias associating science with men, so you respond that the female is more likely to be a non-scientist. You have access to no other information about the individual that can guide your judgement. In such a situation, there will be epistemic benefits to being influenced by the implicit bias: as a result of doing so you will be more likely to make an accurate judgement than if you did not make the association and you will be no less likely to respond appropriately to evidence about the specific individual because there is no such evidence available. As the association tracks something of social reality (say you are in the UK where only 13% people working in the sciences in 2014 were women (WISE 2015)), it will dispose you to respond in certain ways that could, under conditions of this sort, increase your chance of making a correct judgement. Given that there can be cases like these, where there are epistemic benefits to automatically associating members of some social groups more strongly than others with certain traits, there can be corresponding epistemic costs to responding in an egalitarian way, and not making an association of this sort (or not allowing the association to be activated). The dispositions to respond in ways that could increase one's chance of obtaining true beliefs or responding appropriately to the evidence, which come with making the association, will be absent or not manifest.

4. Epistemic costs of implicit stereotyping

Now let us turn to the epistemic costs of automatically associating members of some social groups more strongly than others with certain traits. Why should we think that making

associations that reflect social realities is epistemically costly? To provide an answer to this question, this section describes downstream epistemic costs that can result from engaging in this form of implicit stereotyping: showing how automatic stereotyping leads one to form beliefs that are insensitive to relevant evidence and unlikely to track the truth. The purpose of the outline is to highlight: (i) the vast quantity of downstream epistemic costs that can follow from stereotyping, and, (ii) that epistemic costs follow from stereotyping regardless of whether or not the stereotype reflects some aspect of social reality.[5]

Distortion of memories

The first epistemic cost of automatic stereotyping is that it distorts our memories about individuals to whom we apply a stereotype. When a person applies a stereotype to a particular individual they attend to and remember features of the individual that fit the stereotype better than features that do not fit the stereotype (Rothbart, Evans, and Fulero 1979; Cohen 1981; Srull, Lichtenstein, and Rothbart 1985). For example, Cohen (1981) undertook a study in which participants were shown a clip in which a woman and a man were having a discussion. Some participants were led to believe that the woman was a waitress and others were led to believe that she was a librarian. Although they were shown identical tapes, participants who believed that the woman was a librarian were more likely to remember that she was wearing glasses, and participants who believed that she was a waitress were more likely to remember that she was drinking beer (she was both wearing glasses and drinking beer). In other words, participants were more likely to remember the woman as possessing features that fit the stereotype associated with her job than features that do not. In another study, aimed at studying the effects of implicit racial bias on memory, participants were presented with a description of an ambiguous confrontation in a bar (Levinson 2007). The race of a key protagonist was manipulated: the same description was presented to different participants but in some descriptions the key protagonist was a Caucasian person, William, while in others it was Tyronne, an African American or Kawika, a Hawaiian man. Participants were asked to state whether certain descriptions of features of the scenario were true or false. The study revealed biased memory effects: people recalled the protagonist as having engaged in significantly more aggressive actions when he was named Tyronne than when his name was William or Kawika, regardless of whether or not the actions had really occurred. They were more likely to (truly or falsely) recall an aggressive act if the protagonist was given an African American name rather than a Caucasian or Hawaiian name. Meanwhile, participants were more likely to falsely recall mitigating factors when the person described in the scenario was called Kawika. The memory effects did not correlate with measures of explicit bias, suggesting that implicit biases were responsible for the memory biases. The full extent of the epistemic costliness of memory biases like these is outlined in this section.

What we remember often determines what we believe and the types of information that we draw upon when we make judgements (see, e.g. Rothbart 1981). This means that stereotypes can lead us to form distorted beliefs and make distorted judgements: ones that fail to reflect the true nature of individuals. If one interacts with a black male, and applies the black male stereotype, he might display many characteristics – he might be articulate, knowledgeable and generous – but if he shows any signs of being aggressive then one is likely to remember this better than the other features displayed because aggression is a part of the stereotype of black males (Devine and Elliott 1995). One will then believe and judge the man to be aggressive rather than as possessing positive characteristics. If one interacts with a woman one knows to be a successful career women, so when one

thinks of her the career women stereotype is automatically activated, she might be smart and funny, but if she shows any signs of being cold rather than warm, the coldness will be remembered better than the other features because it fits with the stereotype of career women as competent but cold (Fiske et al. 1999). Rather than reflecting the multifaceted nature of these individuals in one's beliefs one will believe them to possess the stereotypical features that they appear to possess but not the other features. This will be epistemically costly because it will lead to a distorted impression of individuals that fails to accurately reflect the available evidence about them.

Notably, the Cohen study has been interpreted as providing reason for thinking that having information about the social category membership of an individual, in this case information about a person's job, can increase the amount of information that is remembered about that individual (Jussim 2012).[6] Participants in the experiment who were given information about the woman's job before they were shown the tape remembered on average remembered 7% more of the target's traits than participants who were given the same information after watching the tape. It might seem as if having the social category information, and therefore having the stereotype activated, is better from an epistemic perspective than not doing so because it allows more information about the specific case to be gathered. However, although more information is remembered as a result of the possession of the social category information, the information that is remembered is biased so that she is represented as more stereotypical than she really is. Given the choice to have more but biased information, or less but unbiased information, the best choice from an epistemic perspective will often be the unbiased sample.

To see this, consider the following example. A panel is hiring a researcher to work in a neuroscience lab. The panel members sift through the application forms of the candidates before creating a shortlist. Based on the empirical findings just discussed, if the panel members have social category information about the candidates (e.g. their gender, ethnicity, religion), they will remember more of the characteristics of the candidates but they will remember the characteristics in a biased way. So, for example, if a female candidate has taken a career break for maternity and spent a semester at a prestigious foreign institution as a visiting research fellow, the former is likely to be attended to and remembered but not the latter because the maternity break fits the stereotype associating women with the domestic sphere more strongly than careers. The maternity break is likely to be taken to provide evidence of the candidate's lack of commitment to her career and this impression will not be counterbalanced by the information about her prestigious research visit. The result will be a distorted picture of the female's commitment to her career. In a case of this sort individuals would be highly likely to form a more accurate evaluation if they remembered less information, for example, only the candidates' qualifications, but the information was balanced and not biased to fit stereotypes.

This case illustrates how automatic stereotyping can produce a distorted picture of reality even where a stereotype reflects social reality in some respect. While it is true that women are vastly underrepresented in the sciences, and therefore any randomly selected committed scientist is likely to be a man, a stereotype that associates women but not men with being committed to a scientific career can still lead to a distorted picture of females to whom the stereotype is applied.

Misinterpretation of ambiguous evidence

The second epistemic cost of stereotyping is that it leads to the misinterpretation of ambiguous actions. When a piece of behaviour could be interpreted as properly belonging to various

different categories (e.g. aggressive or playful), the activation of a stereotype sets expectancies, leading the ambiguous action to be interpreted in a way that is consistent with the stereotype (Duncan 1976; Sagar and Schofield 1980). For example, in a study undertaken by Duncan (1976) participants were shown a clip of one person shoving another. When the protagonist was black and the victim was white the majority of participants (75%) judged the shove to be an act of violence. When the protagonist was white and the victim was black only a small minority of participants (17%) judged the shove to be an act of violence. The best explanation of the participant's interpretations of the shoving incident is that their judgements were determined by the stereotype associating black people and not white people with violence. The same action was interpreted differently depending on the stereotype that was applied. Devine (1989) found a similar effect following from implicit stereotyping. She found that individuals unconsciously primed with words associated with the stereotype of black people viewed ambiguous behaviour to be more aggressive than those who were not primed in this way. When the automatically activated stereotype associating black people with violence was applied, an inaccurate judgement was made: ambiguous evidence was interpreted in a way that led people to think that the act was violent rather than ambiguous (for further evidence of the influence of implicit associations on the interpretation of ambiguous behaviour see Gawronski, Geschke, and Banse 2003). It is epistemically costly that our stereotypes work in this way because they lead to misperceptions of behavioural evidence.

It is clear that stereotypes can have this epistemic cost while reflecting something of a social reality. Imagine, for example, that a woman displays errors in her speech when explaining a scientific concept at a conference. The speech errors are ambiguous; they are consistent with her lacking knowledge about the topic about which she is speaking, but also with her lacking confidence speaking in a public arena. Because she is a woman, and science and scientific expertise are associated with men, a person applying the scientist stereotype is likely to interpret her ambiguous behaviour as indicative of a lack of knowledge rather than a lack of confidence. On the other hand, a man displaying the same speech errors is likely to be interpreted as knowledgeable but lacking confidence because science, and scientific expertise, is associated with men rather than women. Once again, as women are underrepresented in science, most scientific experts are men, so the stereotype reflects something of social reality. Nonetheless, it can lead to inappropriate interpretations of individuals' behaviours. While the behaviours should be interpreted as ambiguous, they are interpreted in ways fitting with a stereotype.

Failure to notice differences between individuals

A third epistemic cost of automatic stereotyping follows from the way that stereotyping leads individuals to be viewed as a part of a majority or minority group. When people are categorised together due to their traits (e.g. skin colour, gender, financial situation) because a stereotype is activated they become viewed as group members. The group may be a minority or a majority group. Members of minority groups are seen as less diverse, more likely to share characteristics (Bartsch and Judd 1993). For example, male nurses, who are a minority group, are viewed as more similar to each other than female nurses (Hewstone, Crisp, and Rhiannon 2011). Members of minority groups are assumed to share characteristics with other members of their group who have been previously encountered. Differences between members of minority groups are less readily noticed than differences between members of majority groups. This is epistemically costly because it means that details about individuals, which are likely to be relevant to judgements that are made about those individuals, go unnoticed.

From the example of male nurses it is possible to see how even a stereotype that accurately reflects some aspect of social reality can have the results just described. If nursing is stereotyped as female, this reflects the social reality that the majority of nurses are female. A consequence of this will be that the status of males as members of a minority group is made salient. If they are viewed as members of a minority group then they will be viewed as homogenous. People who assess them will be less sensitive to their individual traits, skills and experiences as they will be assumed to resemble previously encountered male nurses. Applying the same principle to the case of women in the sciences, the distinctive characteristics of female scientists are less likely to be noticed than those of their male counterparts, because others will assume that they are similar to previously encountered members of their minority group.

Failure to notice similarities between members of different groups

A related phenomenon is that where individuals are viewed as forming distinct groups due to the characteristics that they possess, in other words, as a result of the activation of a stereotype, differences between the groups are magnified (e.g. Tajfel 1981). A stereotype associating one set of individuals with a characteristic can highlight differences between that set of individuals and others, preventing similarities between members of the different groups from being noticed. Epistemic errors can follow because inaccurate assessments are made of the traits possessed by group members; they are inaccurately assessed as less similar to each other than they really are.

To see how this phenomenon can occur even where a stereotype accurately reflects some aspect of social reality, we can return to the science case. The stereotype associating science with men can lead to male and female scientists being treated as distinct groups. Application of the stereotype can highlight differences between male and female scientists. This can lead people to magnify the differences and fail to see the similarities between male and female counterparts. Where one gender is associated with certain traits, like being committed or uncommitted to their careers, and the differences between males and females are magnified, evidence of those traits are less likely to be noticed when they are present in the other gender.

Failure to track truth in explanations of behaviour

A further epistemic cost to stereotyping is that people who engage in the practice fail to develop explanations of other people's behaviour that are well-grounded in evidence or likely to track the truth about the behaviour's cause or motivation. People who rely on stereotypes tend to explain actions in terms of the dispositions of the actor if the actions fit the stereotype of the actor's social group, but they explain the actions in terms of the situation of the actor rather than the disposition if the actions fail to fit the stereotype. For example, if a black person shoves a white person, their action is attributed to their dispositions, but if a white person shoves a white person, the action is explained in terms of features of the situation (Duncan 1976). The same action, produced by the same cause in an extremely similar situation, is explained in different ways depending on the social group membership of the individual being assessed. In addition to this, stereotypes are often taken to provide sufficient explanations of individuals' behaviours so further explanations are not sought where they should be (Sanbonmatsu, Akimoto, and Gibson 1994). In the science case, speech errors displayed by a female while explaining a scientific concept might be explained by lack of scientific expertise if the stereotype associating scientific

expertise with men is activated, where another explanation would be far more adequate. Even where there could be some sense in which the stereotype could be said to reflect social realities, the error of taking someone's social group membership to provide an explanation of their behaviour rather than seeking further relevant information could occur and be epistemically costly, leading to a lack of true understanding.

Inappropriate associations and cognitive depletion

It might be thought that when people engage in automatic stereotyping that reflects the social reality at least they make appropriate associations: If people automatically associate science with men or women with nurturing then the association could be appropriate because of reflecting social reality. However, currently popular models of stereotypes and stereotyping provide good reason to think that when people make associations of this type they consequently either make other associations that fail to reflect the social reality (Kelly and Roedder 2008) or fail to make other associations that would reflect social reality.

On the associative network view, stereotypes are a set of linked attributes (Manis, Nelson, and Shedler 1988; Carlston 1992). The stereotype of a scientist is a set of attributes – for example, lab-coat wearing, bespectacled, male, hardworking, conscientious, expert, attentive to detail, innovative thinker – which are linked together in the mind of the believer. The perception of one of the attributes leads to the activation of associated concepts. Not only are concepts activated, however, feelings of positivity or negativity can also be activated. On this model, stereotyping involves a number of different associations, which are like a tightly knit cluster of beliefs. When one holds a cluster of beliefs, one belief might be true but it might trigger a number of false beliefs. The same is true of associations. One association might reflect social reality in some respect but trigger other associations that do not reflect social reality. For example, the association between science and men might trigger a further association between men and attentiveness to detail and high IQs. Individual men might therefore be associated with these positive characteristics even where the positive associations are inappropriate. Men might be associated with attentiveness to detail or a high IQ when they are no more likely to have these traits than women. Meanwhile, the associations might not be made between women and these positive characteristics because the stereotype does not associate them with science.

The prototype model of stereotypes also suggests that inappropriate associations will be made and appropriate associations left unmade. On the prototype model, stereotypes are abstract representations of typical features of members of a group (Cantor and Mischel 1979), for example, the stereotype of a bird might be an abstract representation of a small-winged creature, creature with beak, flying-creature, creature that tweets. Whether a stereotype is applied to any individual depends on the similarity between traits of the individual and a set of typical features ordinary associated with members of the group. But the traits of the individual that determine whether the stereotype is applied are usually superficial features of the individual, such as signs from their face of their gender or race (see, e.g. Ashmore and Del Boca 1979; Bodenhausen and Macrae 1998). These features of human psychology combine to mean that the associations are likely to often be applied inappropriately. A case in which an association would be made when it shouldn't be is the following: a woman displays superficial features that are a part of the abstract representation of a wife and mother, that is, feminine facial and bodily features. Other features of the abstract representation of a wife and mother are triggered. However, the woman has few other characteristics associated with being a wife and mother: most notably, she has no

husband or children. The same case could be one in which associations are not made but the should be. The woman displays superficial features that are not associated with science, that is, the same feminine facial and bodily features. Other features that are a part of the abstract representation of scientists are therefore not associated with her. The perception of the superficial signs that she is a woman inhibits the scientist stereotype. However, the individual possesses many other features associated with the scientist stereotype: she has expertise in science and related positive traits like attentiveness to detail or innovative thinking.

Things seem little better on the exemplar view of stereotypes (see, e.g. Smith and Zárate 1992). According to this view, stereotypes represent groups through particular concrete examples, for instance, the stereotype of a bird is represented through a robin. Whether or not a stereotype is activated in a particular case is determined by the similarity between the target individual and the exemplar. The fact remains that superficial features of an individual determine whether they are judged to be similar to the exemplar. This means that a cluster of features possessed by a particular exemplar will be associated with a target individual only if they possess superficial features that are similar to those of the exemplar. For example, Einstein might be the exemplar of a scientist. Then a cluster of features associated with Einstein (expertise in science, genius, innovative thinking) will be automatically associated with an individual only if they display superficial features associated with him (being male, looking eccentric, etc.). Individuals who do not display these superficial characteristics will not be stereotyped as a scientist, and will therefore not have Einstein's features automatically associated with them. This is bad news for many women, who are much less likely than their male counterparts to possess at least one of the superficial features of exemplars of science like Einstein, that is, having masculine facial features. It is also bad news for those of us who want to make accurate judgements about individual women who are scientists but do not share the superficial features of familiar exemplars of scientists because if the exemplar view is correct we are unlikely to automatically associate these individuals with features that they actually possess (expertise in science, attentiveness to detail, innovative thinking). In addition to this, individuals who share superficial characteristics with exemplars of scientists are likely to be automatically associated with other characteristics possessed by those exemplars even if the only features that they share are the superficial ones (e.g. their facial features).

Finally, stereotyping claims are sometimes argued to be generic claims, with a similar structure to claims like "ducks lay eggs" or "mosquitoes carry the West Nile virus" (see, e.g. Leslie 2007). They are not universal claims, but they suggest that members of a group typically, characteristically or strikingly possess certain features. Sally Haslanger (2010) presents reason for thinking that if stereotypical thought has the structure of generic claims, it can lead to inappropriate associations, even in a context in which the stereotypes applied reflect something of social reality. Haslanger points out that generic claims that involve associations between members of certain social groups and particular characteristics can imply that there is something essential about the nature of the group members in virtue of which they possess the characteristics associated with them. These essentializing judgements can be inaccurate because it is really the social context in which the group members are located that means that they are more likely than others to possess the traits associated with them. For example, women might not only be associated with being nurturing but also with having a nature in virtue of which they are disposed to be nurturing. While the association between women and nurturing might reflect something of reality, as a result of social constructs that lead women to occupy a nurturing role, it can lead to a further association that fails to reflect reality: the association of women with an essentially nurturing nature.

A number of negative epistemic consequences can follow from making inappropriate associations or failing to make appropriate associations. We can make poor inferences, assuming that people have characteristics that they lack, or that they do not have characteristics that they do have (Egan 2011). For example, we can assume that women are lacking in scientific expertise but nurturing, because of the superficial characteristics they display. Alternatively, we can notice that we are making erroneous judgements and expend a large amount of cognitive energy to suppress our stereotyping responses (Egan 2011; Gendler 2011). This can lead to the depletion of our cognitive resources, which can in turn cause errors of judgement. The latter phenomenon has been illustrated through experiments in which white participants are required to interact with black experimenters before completing a cognitive task (Richeson and Shelton 2007; cited in Gendler 2011). Those who were found to produce high scores on implicit measures of racial stereotyping, indicating that they make strong negative implicit associations with black people, performed poorly on the cognitive tasks, seemingly because their cognitive resources were depleted by efforts to suppress their stereotyping.[7]

5. Epistemic benefits of failing to reflect social realities

Now let us consider how the epistemic costs outlined in section 4 can be avoided by our cognitions failing to reflect social realities and us thereby making more egalitarian automatic responses. I focus on two of the strategies that were outlined in the introduction: making curriculum vitae anonymous in recruitment processes and considering counter-stereotypical examples. Both strategies involve reducing the chance of social realities being reflected in judgements by preventing stereotypes from being automatically triggered. The first strategy involves an individual being placed in a position in which they cannot reflect certain social realities in their judgements. If one does not have access to information about the social category to which an individual belongs then one cannot make automatic associations with an individual due to their group membership. The second strategy involves choosing for social realities to be misrepresented. It involves putting in place a situation in which people respond as if the social reality is different to how it is. It leads to an egalitarian response; people responding as if men and women are represented equally within the sciences, or rates of involvement in crime are equal across different races. As stereotyping involves associating particular social groups more than others with traits the egalitarian response will not involve stereotyping. Both strategies therefore avoid the epistemic costs outlined in section 4 by avoiding instances of automatic stereotyping.

How, more specifically, are each of the costs avoided? First, if either of the strategies is adopted then information about individuals that fits the stereotype of their social group is no more likely to be remembered than information that does not, because a stereotype will not be triggered. Second, for the same reason, ambiguous evidence about individuals, for example, actions that could potentially be viewed as signs of ignorance or lack of confidence, will not be misinterpreted in a way that is fitting with the stereotype of their social group. Third, individuals are less likely to be assumed to share traits with previously encountered members of their social group if their group is a minority group (e.g. women in science, an ethnic minority group). If the first strategy is adopted then the individual's status as a member of a minority social group will not influence the judgements made about them because it will not be known. If the second strategy is adopted then sets of individuals will be less likely to be divided into separate groups on the basis that they are viewed as possessing different traits because they will be viewed as equally likely to possess certain traits, so

they will not be separated into minority and majority groups. Minority groups will not be assumed to share the same characteristics, because they will not be viewed as separate from majority groups, so people will not fail to notice differences between individuals. Fourth, and relatedly, differences between an individual being assessed and members of other social groups will not be magnified: either because their social group membership is not known or because they will not be treated as a member of a distinct social group. Fifth, as stereotypes will not be triggered if either strategy is adopted, behaviours that are fitting with the stereotype will not be explained in terms of the dispositions of the individual while behaviours that are counter-stereotypical will not be explained in terms of the individual's situation. Explanations of individuals and their behaviours will not focus on stereotypes where they should focus on other characteristics. Sixth, and finally, as a stereotype is not triggered if either strategy is adopted it is less likely that the believer will make associations that are inaccurate and not endorsed. Poor inferences and the cognitive depletion that can accompany the suppression of disavowed stereotyping are therefore unlikely to follow.

This discussion provides good reason for thinking that cognitions that fail to reflect social realities can meet the *epistemic benefit* condition. Although they bring the epistemic cost of preventing the agent from being disposed to respond in some ways that will, under some circumstances, increase her chances of obtaining true beliefs or responding appropriately to the evidence, they bring substantial epistemic benefits by leading to the avoidance of the significant costs that are associated with stereotyping even when the stereotypes applied reflect some aspect of social reality.

6. No alternatives

The same cognitions can also meet the *no alternatives* condition because alternative cognitions that would confer the same epistemic benefits without the epistemic costs are not available to the agent at the time. Given that the cognitions that are the focus of the current discussion are automatic associations that bring epistemic costs because they fail to reflect social realities, the relevant alternative cognitions are automatic cognitions that reflect the social realities (e.g. cognitions that associate science with men or women with being nurturing). For the cognition that fails to reflect social reality to be epistemically innocent, it must be the case that the believer could not have an automatic cognition that reflects the social reality while also gaining the epistemic benefits that follow from failure to do so.

There is good reason to think that this is often the case. The psychological literature discussed in the previous section suggests that where people are sensitive to differences between members of different social groups, thereby having a cognition that is non-faulty in the sense that it reflects social realities, they are highly likely to engage in automatic stereotyping (see also Gendler 2011) that leads to the sorts of errors outlined in section 4. There are consequently often two forms of cognition available to the believer in the face of a social inequity: (i) a cognition that fails to reflect the social realities and avoids the epistemic costs associated with stereotyping, or (ii) a cognition that reflects the social realities but fails to avoid the epistemic costs associated with stereotyping. There is no alternative cognition available to the thinker that reflects the social realities but provides the benefits associated with avoiding the costs outlined in (4a–f).

7. The lesser of two epistemic evils

Finally, there is good reason for thinking that automatic cognitions that fail to reflect social realities can belong to the subset of epistemically innocent cognitions that are the lesser of

two epistemic evils. There is a weaker and a stronger claim that can be made in this vicinity.[8]

Weaker lesser evil claim: there are *some* conditions under which cognitions that fail to reflect social realities are the lesser of two epistemic evils.

Stronger lesser evil claim: there are *many* conditions under which cognitions that fail to reflect social realities are the lesser of two epistemic evils.

Substantiation of the weaker claim is all that is required to put pressure on the inference from evidence that a stereotype reflects some aspect of social reality to the conclusion that it is best from an epistemic perspective to make the association encoded in the stereotype. If there are some conditions where the best epistemic consequences follow if one does not reflect social realities, evidence that a stereotype reflects some aspect of social reality does not establish that it is best from an epistemic perspective to apply it.

The following example illustrates the correctness of the weaker claim. Imagine a person making an assessment of the contribution that a female scientist could make to their research team. If they automatically respond by applying the stereotype associating science with men (the way that reflects the social reality that women are underrepresented in the sciences), then the disposition that they manifest will be one that would increase the chance of them making an accurate assessment of the probability that a randomly selected individual scientist is a woman or a man, or a randomly selected individual person is a scientist, in the absence of informative case-specific information. However, if they respond in a way that does not involve stereotyping (the egalitarian response that fails to reflect the social reality), then they will respond in a way that brings the advantages associated with avoiding the following numerous pitfalls: (i) memory distortions that would make them remember features of the candidate that fit the stereotype that scientific experts are men better than other features; (ii) viewing ambiguous behaviours of the candidate as evidence of lack of expertise; (iii) failing to notice differences between the candidate and other, previously encountered female scientists; (iv) failing to notice similarities between the candidate and male scientists who have positive features that suggest that they are experts; (v) the tendency to assume that any behaviours that are stereotypical of non-experts (e.g. lack of confidence speaking about the subject) are indicative of the dispositions of the candidate rather than the situation that she is placed in; and, finally, (vi) the tendency to make all sorts of associations with the candidate that are inaccurate and not endorsed, leading cognitive efforts to be expended on suppressing the stereotyping rather than, for example, concentrating on the evidence relating to the levels of expertise of the candidate. The benefits associated with avoiding these pitfalls in this context vastly outweigh the benefits that come from automatically responding in a way that reflects the statistical reality that women are underrepresented in the sciences.

It might be tempting to draw the universal conclusion here that the epistemic benefits of failing to reflect social realities by not engaging in automatic stereotyping always outweigh the epistemic costs. This would be too swift. Sometimes individuating information about the specific case is sparse and therefore, with information about social group membership as the main guide of one's judgements, one might be more likely to obtain true beliefs if one makes an association that reflects the social reality. However, cases of this sort are few and far between. More often, the accuracy of our judgements depends on our ability to access and process case-specific information, for example, information about particular individuals who might be scientists, their past experience, their qualifications, their analytic skills, etc. What the discussion in this paper shows is that when information of this sort is available, automatic associations reflecting social realities are likely to prevent us from appropriately accessing and processing the information, leading to the types of distortion of case-specific

information outlined in section 4. As we *often* have informative, case-specific information available to us, which can be distorted via automatic stereotyping, it will *often* be best from an epistemic perspective not to make automatic associations. Failing to reflect background information about social inequalities, so that we respond in egalitarian ways, will consequently *often* increase our chances of making accurate judgements. Here, then, we find support for *the stronger lesser evil claim*. Not only is it sometimes the case that cognitions that fail to reflect social realities are the lesser of two epistemic evils, it is often the case.

8. Objections and replies

It might be objected that sometimes it is not possible to get a proper understanding of the situation of an individual or set of individuals without awareness of certain facts about the social realities that they have encountered in their everyday lives (Madva 2016a). To properly understand the difficulties faced by women in science, such as the lack of mentoring opportunities, it is necessary to comprehend the social reality of female underrepresentation in the sciences. In addition to this, it might be argued that if we are not aware of social inequalities, we will not be able to take measures to improve our epistemic situation.[9] If we are not aware of the underrepresentation of women in the sciences, we will not realise that we are likely to be disposed to respond in a biased and distorted way to female scientists because they are members of a minority group. It might therefore seem to be necessary to accurately reflect social realities in our judgements in order to gain the epistemic good of understanding and in order to recognise the need to mitigate the negative effects of bias and prejudice.

This objection does provide reason to be cautious in the adoption of some strategies that might be used to mitigate bias and prejudice. Take, for example, strategies to change the associations that people automatically make to ensure that they are more egalitarian. The arguments presented in this paper provide reason for thinking that in some situations these strategies will bring overall epistemic gains by reducing the chance of stereotyping. However, if, to achieve the egalitarian aim, one were exposed to a distorted picture of the situation faced by some individuals within society, then the strategy could bring substantial epistemic costs. If, for example, one is exposed to a portrayal of science in the media in which women and men were represented equally in the sciences, and supplied with no other information about the problems faced by women as an underrepresented group within science, then one would be unlikely to have an adequate understanding of the plight of individual women scientists, including the bias and prejudice they might face. Caution should therefore be urged with regards to strategies of this sort: although they can bring the epistemic gains associated with failing to reflect social realities by responding in the egalitarian way, unless they are adopted with care they can bring substantial epistemic costs.

However, the same objections can also be taken to illustrate a way to maximise one's epistemic gains while failing to reflect social realities in one's automatic responses. They suggest that the best way to achieve epistemic goals can be to achieve two things simultaneously: (i) having an explicit understanding of social inequalities; and (ii) avoiding the costs of automatic stereotyping by facilitating egalitarian automatic responses.[10] By achieving both of these goals the agent could ensure that they obtain the benefits of failing to reflect social inequalities in their automatic responses while nonetheless achieving understanding of the plight faced by members of some social groups and the way that the inequalities can lead to bias and prejudice.

The achievement of both of these goals simultaneously is not without complication. There is reason to think that mere awareness of social inequalities can be enough to

trigger automatic associations and implicit stereotyping (Correll et al. 2002; see also Gendler 2011; Madva 2016b for discussion). This means that, for example, the awareness of the underrepresentation of women in the sciences required to understand the difficulties faced by female scientists and the way that one might respond in a biased way which needs to mitigated, can be enough to trigger an association between science and men, leading to automatic stereotyping. But while the simultaneous achievement of both of these goals might be difficult, it is not impossible – there is good reason to think that strategies can be adopted to achieve this end.

To see this point, we can return to the example of Anita, who is on a hiring panel, recruiting a scientific researcher. Let us assume that she is aware of the social reality that women are underrepresented in the sciences, that they might consequently have suffered difficulties not encountered by male counterparts, and that she is likely to be biased in judgements she makes about candidates because of associating science with men. She has an explicit understanding of the relevant social inequality and the need to take action to mitigate its negative effects. Her awareness of this need to take action does not preclude her from responding in an egalitarian way when sifting through a large pile of curriculum vitae of job candidates because she ensures that CVs are anonymous when the sifting process is undertaken. Her automatic responses are therefore egalitarian and she avoids the epistemic costs of automatically associating scientific expertise with men that have been outlined in this paper. In addition to this, she can ensure that her awareness of the underrepresentation of women in the sciences is manifest in an understanding of the difficulties that they might have faced: Additional time and attention might be devoted to the application materials of female candidates, on the basis that they might have suffered difficulties not faced by male candidates. If this process is undertaken with due care and attention it will be possible to avoid it being influenced by automatic stereotyping because this form of stereotyping can be reduced by decreasing time pressures and enabling evaluators to fully attend to the task at hand. For instance, Martell (1991) asked participants to evaluate descriptions of work undertaken by male and female police officers. The work of males was evaluated more favourably when the evaluation was made under time pressure and with competing demands on the evaluators' attention. However, when time pressures were eliminated and evaluators were able to attend fully to the task, the discrepant judgements were abated and automatic stereotyping was reduced (for discussion see Saul 2012).

While the objections considered in this section provide reason to re-consider some strategies that might be adopted to overcome the negative effects of automatic stereotyping, they do not therefore show that it cannot be best from an epistemic perspective to fail to reflect social inequalities in our automatic responses. They show instead that we can maximise the epistemic benefits we gain by combining an understanding of the situations faced by individuals in virtue of their social group membership, an understanding of the need to moderate our responses to those individuals, and automatic responses that fail to reflect social inequalities. Exemplary strategies to mitigate prejudice and bias can facilitate the achievement of each of these goals simultaneously.

Conclusion

The main message to be taken from the current discussion is that often the ethical option, reacting in the egalitarian way that fails to reflect social realities such as the underrepresentation of women in the sciences or the higher number of women occupying nurturing roles in our automatic responses, is also best from an epistemic perspective. It is the epistemically innocent option and the lesser of two epistemic evils: although it brings epistemic costs,

these are often outweighed by the epistemic benefits. Even where an automatically activated stereotype reflects some aspect of social reality, the activation of the stereotype can be epistemically costly, and it can be best from an epistemic perspective for the stereotype not to be activated. This argument has important implications for how to respond to a natural defence of automatic stereotyping: that if a stereotype reflects some aspect of social reality, by applying the stereotype one is following the facts in a way that will increase the chance of one making accurate judgements. The argument presented in this paper suggests that this defence is often wrong-headed. Often it will be best from an epistemic perspective not to reflect social reality in one's automatic responses because reflecting reality will increase the chance that one will engage in automatic stereotyping, suffering serious downstream epistemic costs.

Acknowledgements

Thanks go to Lisa Bortolotti, Ema Sullivan-Bissett, Jules Holroyd, Sophie Stammers, audiences at the *False but Useful Beliefs* conference, the Early Career Mind Network meeting at University of Warwick, and the Implicit Bias Reading Group at the University of Birmingham, as well as two anonymous referees for comments on versions of this paper.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

The author acknowledges the support of the European Research Council under the Consolidator grant agreement number [616358] for a project called *Pragmatic and Epistemic Role of Factually Erroneous Cognitions and Thoughts* (PERFECT)

Notes

1. There is an important feature of this example: egalitarian principles come into conflict with evaluatively loaded automatic responses (e.g. women are better at childrearing than men) that are fitting with reality. For further discussion of examples of this sort see Haslanger (2010, 2015). For current purposes, evaluatively loaded automatic responses are treated similarly to less evaluatively loaded responses (e.g. scientists are men) because the important point raised by this paper is that the automatic associations manifest in implicit bias bring epistemic costs, and the vast majority of these costs will follow regardless of whether they are evaluatively loaded. That being said, it is worth noting that it is likely that there will be additional epistemic costs that follow exclusively from evaluatively loaded automatic responses. Thanks go to an anonymous referee and Sophie Stammers for help developing this point.
2. The frequency with which the epistemic costs of stereotyping outweigh the epistemic benefits is obviously an empirical issue. I aim to show that there is good reason to think that the psychological findings present good reason for thinking that the empirical reality is that the costs often outweigh the benefits because there are many ways and conditions under which the costs can manifest and few by which benefits can.
3. There is some debate about whether implicit biases are beliefs (Mandelbaum 2015), belief-like states (Schwitzgebel 2010; Levy 2015) or belong in a distinct psychological category from beliefs (Gendler 2008a, 2008b, 2011; Madva 2016a). There is also debate over whether they are propositional (Mandelbaum 2015) or merely associative (Gendler 2008a, 2008b). This paper remains neutral with regards to these issues, all that matters for the current discussion is that implicit bias involves the automatic association between social groups and their members and certain traits and evaluative content. Whether the association comes in the form of a belief or a psychological state with propositional structure does not matter. My focus is on the epistemic costs and benefits of these associations.

4. See section 8 for how some of these strategies can be better than others from an epistemic perspective, depending on the context in which they are adopted.
5. Gendler (2011) outlines three epistemic costs that can follow from implicit bias. One of these costs is outlined in section 4f. This is the only one of Gendler's costs to be included in this discussion because Mugg (2013) has convincingly argued that the other two costs are not epistemic costs for the person engaging in the stereotyping.
6. Something similar might be said about the Levinson study: people remembered more aggressive actions that actually occurred when they harboured a negative implicit bias against Black people. It might therefore be argued that the implicit bias is epistemically beneficial because it leads to the retention of some negative information that was not remembered by people who were not biased.
7. Joshua Mugg (2013) has argued that there are long-term gains associated with using cognitive resources to suppress stereotyping. He makes a comparison between the use of executive function to control stereotyping and the use of executive function by a bilingual suppressing their non-active language. Over time the bilingual gains from suppression of the non-active language because her executive functioning is strengthened. Mugg's suggestion is no threat to the claims made in the current discussion. The identification of long-term gains does not undermine the claim crucial to this paper, that there can be epistemic costs at a given time to taking action to prevent stereotyping.
8. Thanks to Ema Sullivan-Bissett for encouragement to clarify this distinction.
9. Thanks to Lisa Bortolotti for encouraging me to consider this point.
10. See Madva (2016b) for discussion of a similar approach to implicit bias, according to which people should aim to control the accessibility of information about social realities.

Notes on contributor

Katherine Puddifoot is a research fellow at the University of Birmingham, United Kingdom, where she works on the ERC-funded 'Project PERFECT'. She was previously a teaching fellow at the University of Glasgow and the University of Bristol. She received her PhD in philosophy from the University of Sheffield with a dissertation on epistemic naturalism. Her research fields are philosophy of psychology, epistemology, and philosophy of medicine.

References

Ashmore, R. D., and F. K. Del Boca. 1979. "Sex Stereotypes and Implicit Personality Theory: Toward a Cognitive-Social Psychological Conceptualization." *Sex Roles* 5 (2): 219–248.

Bartsch, Robert A., and Charles M. Judd. 1993. "Majority-Minority Status and Perceived Ingroup Variability Revisited." *European Journal of Social Psychology* 23: 471–483.

Beeghly, Erin. 2015. "What Is a Stereotype? What Is Stereotyping?" *Hypatia* 30 (4): 675–691.

Blair, I. V., J. E. Ma, and A. P. Lenton. 2001. "Imagining Stereotypes Away: The Moderation of Implicit Stereotypes Through Mental Imagery." *Journal of Personality and Social Psychology* 81 (5): 828–841.

Blum, Lawrence. 2004. "Stereotypes and Stereotyping: A Moral Analysis." *Philosophical Papers* 33 (3): 251–289.

Bodenhausen, Galen V., and C. N. Macrae. 1998. "Stereotype Activation and Inhibition" In *Stereotype Activation and Inhibition: Advances in Social Cognition Vol. 11*, edited by R. S. Wyer, 1–52. Mahwah, NJ: Erlbaum.

Bortolotti, Lisa. 2015a. "Epistemic Benefits of Elaborated and Systematized Delusions in Schizophrenia." *British Journal for the Philosophy of Science*. doi:10.1093/bjps/axv024.

Bortolotti, Lisa. 2015b. "The Epistemic Innocence of Motivated Delusions." *Consciousness and Cognition* 33: 490–499.

Brownstein, Michael. 2015. "Implicit Bias." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Spring ed. http://plato.stanford.edu/archives/spr2015/entries/implicit-bias/.

Cantor, Nancy and Walter Mischel. 1979. "Prototypes in Person Perception." *Advances in Experimental Social Psychology* 12: 3–52.

Carlston, Donal E. 1992. "Impression Formation and the Modular Mind: The Associated Systems Theory." In *The Construction of Social Judgments*, edited by L. Martin and A. Tesser, 301–341. Hillsdale, NJ: Erlbaum.

Cohen, Claudia E. 1981. "Personal Categories and Social Perception: Testing Some Boundaries of the Processing Effects of Prior Knowledge." *Journal of Personality and Social Psychology* 40: 441–452.
Correll, J., B. Park, C. M. Judd, and B. Wittenbrink. 2002. "The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals." *Journal of Personality and Social Psychology* 83 (6): 1314–1329.
Devine, Patricia. 1989. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology* 56 (1): 5–18.
Devine, Patricia G., and Andrew J. Elliott. 1995. "Are Racial Stereotypes Really Fading? The Princeton Trilogy Revisited." *Personality and Social Psychology Bulletin* 21 (11): 1139–1150.
Duncan, Birt L. 1976. "Differential Social Perception and Attribution of Intergroup Violence: Testing the Lower Limits of Stereotyping of Blacks." *Journal of Personality and Social Psychology* 34: 590–598.
Egan, Andy. 2011. "Comments on Gendler, 'On the Epistemic Costs of Implicit Bias'." *Philosophical Studies* 156 (1): 65–79.
Fiske, Susan T., Jun Xu, Amy C. Cuddy, and Peter Glick. 1999. "(Dis)Respecting Versus (Dis)Liking: Status and Interdependence Predict Ambivalent Stereotypes of Competence and Warmth." *Journal of Social Issues* 55 (3): 473–489.
Gawronski, Bertram, Daniel Geschke, and Rainer Banse. 2003. "Implicit Bias in Impression Formation: Associations Influence the Construal of Individuating Information." *European Journal of Social Psychology* 33 (5): 573–589.
Gendler, Tamar Szabó. 2008a. "Alief and Belief." *Journal of Philosophy* 105 (10): 634–663.
Gendler, Tamar Szabó. 2008b. "Alief in Action (and Reaction)." *Mind and Language* 23 (5): 552–585.
Gendler, Tamar Szabó. 2011. "On the Epistemic Costs of Implicit Bias." *Philosophical Studies* 156 (1): 33–63.
Haslanger, Sally. 2010. "Ideology, Generics and the Common Ground." In *Feminist Metaphysics: Essays on the Ontology of Sex, Gender and the Self*, edited by Charlotte Witt, 179–207. Dordrecht: Springer.
Haslanger, Sally. 2015. "Social Structure, Narrative and Explanation." *Canadian Journal of Philosophy*. doi:10.1080/00455091.2015.1019176.
Hewstone, Miles, Richard J. Crisp, and N. Turner Rhiannon. 2011. "Perceptions of Gender Group Variability in Majority and Minority Contexts: Two Field Studies with Nurses and Police Officers." *Social Psychology* 42: 135–143.
Hilton, James L., and William von Hippel. 1996. "Stereotypes." *Annual Review of Psychology* 47: 237–271.
Holroyd, Jules, and Katherine Puddifoot. Forthcoming. "Implicit Bias and Prejudice." In *Routledge Handbook of Social Epistemology*, edited by Miranda Fricker, Peter J. Graham, David Henderson, Nikolaj Pedersen, and Jeremy Wyatt.
Jussim, Lee. 2012. *Social Perception and Social Reality.* Oxford: Oxford University Press.
Kelly, Daniel, and Erica Roedder. 2008. "Racial Cognition and the Ethics of Implicit Bias." *Philosophy Compass* 3 (3): 522–540.
Leslie, Sarah Jane. 2007. "Generics and the Structure of the Mind." *Philosophical Perspectives* 21 (1): 375–403.
Letheby, Christopher. 2016. "The Epistemic Innocence of Psychedelic States." *Consciousness and Cognition* 39: 28–37.
Levinson, Justin D. 2007. "Forgotten Racial Equality: Implicit Bias, Decision making, and Misremembering." *Duke Law Journal* 57: 345–424.
Levy, Neil. 2015. "Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements." *Noûs* 49 (4): 800–823.
Madva, Alex. 2016a. "Why Implicit Attitudes Are (probably) Not Beliefs." *Synthese*. doi:10.1007/s11229-015-0874-2.
Madva, Alex. 2016b. "Virtue, Social Knowledge and Implicit Bias." In *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, edited by Brownstein and Saul, 191–215. Oxford: Oxford University Press.
Mandelbaum, Eric. 2015. "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias." *Nous*. doi:10.1111/nous.12089.

Manis, Melvin, Thomas E. Nelson, and Jonathan Shedler. 1988. "Stereotypes and Social Judgment: Extremity, Assimilation, and Contrast." *Journal of Personality and Social Psychology* 55: 28–36.
Martell, Richard F. 1991. "Sex Bias at Work: The Effects of Attentional and Memory Demands on Performance Ratings of Men and Women." *Journal of Applied Social Psychology* 21 (23): 1939–1960.
Mugg, Joshua. 2013. "What Are the Cognitive Costs of Racism." *Philosophical Studies* 166 (2): 217–229.
Puddifoot, Katherine. forthcoming. "Stereotyping: The Multifactorial View." *Philosophical Topics*.
Richeson, Jennifer A., and J. Nicole Shelton. 2007. "Negotiating Interracial Interactions: Costs, Consequences, and Possibilities." *Current Directions in Psychological Science* 16 (6): 316–320.
Rothbart, M. 1981. "Memory Processes and Social Beliefs." In *Cognitive Processes in Stereotyping and Intergroup Behaviour*, edited by D. L. Hamilton, 145–182. Hillsdale, NJ: Erlbaum.
Rothbart, M., M. Evans, and S. Fulero. 1979. "Recall for Confirming Events: Memory Processes and the Maintenance of Social Stereotypes." *Journal of Experimental Social Psychology* 15: 343–355.
Sagar, H. Andrew, and Janet Ward Schofield. 1980. "Racial and Behavioral Cues in Black and White Children's Perceptions of Ambiguously Aggressive Acts." *Journal of Personality and Social Psychology* 39 (4): 590–598.
Sanbonmatsu, Daniel M., Sharon A. Akimoto, and Bryan D. Gibson. 1994. "Stereotype-Based Blocking in Social Explanation." *Personality and Social Psychology Bulletin* 20: 71–81.
Saul, Jennifer. 2012. "Ranking Exercises in Philosophy and Implicit Bias." *Journal of Social Philosophy* 43: 256–273.
Saul, Jennifer. 2013. "Scepticism and Implicit Bias." *Disputatio* 5 (37): 243–263.
Schwitzgebel, Eric. 2010. "Acting Contrary to Our Professed Beliefs, or the Gulf Between Occurrent Judgment and Dispositional Belief." *Pacific Philosophical Quarterly* 91: 531–553.
Smith, Elliot R., and Michael A. Zárate. 1992. "Exemplar-Based Model of Social Judgment." *Psychological Review* 99 (1): 3–21.
Srull, T. D., M. Lichtenstein, and M. Rothbart. 1985. "Associative Storage and Retrieval Processes in Person Memory." *Journal of Experimental Psychology: Learning, Memory and Cognition* 11: 316–345.
Steinpreis, Rhea E., Katie A. Anders, and Dawn Ritze. 1999. "The Impact of Gender on the Review of the CVs of Job Applicants and Tenure Candidates: A National Empirical Study." *Sex Roles* 41 (718): 509–528.
Stewart, Brandon D., and B. Keith Payne. 2008. "Bringing Automatic Stereotyping Under Control: Implementation Intentions as Efficient Means of Thought Control." *Personality and Social Psychology Bulletin* 34: 1332–1345.
Sullivan-Bissett, Ema. 2015. "Implicit Bias, Confabulation, and Epistemic Innocence." *Consciousness and Cognition* 33: 548–560.
Tajfel, Henri. 1981. *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge: Cambridge University Press.
WISE. 2015. "Women in Science, Technology, Engineering and Mathematics: The Pipeline from Classroom to Boardroom UK Statistics 2014." https://www.wisecampaign.org.uk/uploads/wise/files/WISE_UK_Statistics_2014.pdf.