Implicit Bias, Confabulation, and Epistemic Innocence

Sullivan-Bissett, Ema

DOI:
10.1016/j.concog.2014.10.006

Document Version
Peer reviewed version

Citation for published version (Harvard):
Sullivan-Bissett, E 2015, 'Implicit Bias, Confabulation, and Epistemic Innocence', *Consciousness & Cognition*, vol. 33, pp. 548-560. https://doi.org/10.1016/j.concog.2014.10.006

Consciousness and Cognition 33 (2015) 548–560

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Implicit bias, confabulation, and epistemic innocence ☆

Ema Sullivan-Bissett *

Department of Philosophy, ERI Building, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom

ARTICLE INFO

Article history:
Received 6 June 2014
Revised 13 October 2014
Accepted 14 October 2014
Available online 20 November 2014

Keywords:
Confabulation
Epistemic benefit
Epistemic evaluation
Imperfect cognitions
Implicit bias

ABSTRACT

In this paper I explore the nature of confabulatory explanations of action guided by implicit bias. I claim that such explanations can have significant epistemic benefits in spite of their obvious epistemic costs, and that such benefits are not otherwise obtainable by the subject at the time at which the explanation is offered. I start by outlining the kinds of cases I have in mind, before characterising the phenomenon of confabulation by focusing on a few common features. Then I introduce the notion of *epistemic innocence* to capture the epistemic status of those cognitions which have both obvious epistemic faults and some significant epistemic benefit. A cognition is epistemically innocent if it delivers some epistemic benefit to the subject which would not be attainable otherwise because alternative (less epistemically faulty) cognitions that could deliver the same benefit are unavailable to the subject at that time. I ask whether confabulatory explanations of actions guided by implicit bias have epistemic benefits and whether there are genuine alternatives to forming a confabulatory explanation in the circumstances in which subjects confabulate. On the basis of my analysis of confabulatory explanations of actions guided by implicit bias, I argue that such explanations have the potential for epistemic innocence. I conclude that epistemic evaluation of confabulatory explanations of action guided by implicit bias ought to tell a richer story, one which takes into account the context in which the explanation occurs.

0. Introduction

In this paper I explore the nature of confabulatory explanations of action guided by implicit bias in the non-clinical population. My aim is to highlight the potential epistemic benefits of some confabulatory explanations and tell a richer story about the overall epistemic status of such explanations. Although confabulation is characterised and often even defined on the basis of its epistemic costs, I argue that some confabulations can play a positive role, not just because they act as a psychological defence by enhancing coherence, stability, self-confidence, and well-being (Ramachandran, 1996: 351; Fotopoulou, 2008: 542), but also because they bestow epistemic benefits which are otherwise unavailable.

In section one I describe two imagined cases of confabulatory explanations of actions guided by implicit bias.[1] In section two, I characterise non-clinical confabulation by identifying some common features of the phenomenon, features shared by my two imagined cases. In section three, I introduce the notion of *epistemic innocence* to capture the status of those

☆ This article is part of a special issue of this journal on Costs and Benefits of Imperfect Cognitions.

* Address: ERI Building, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom.
E-mail address: e.l.sullivan-bissett@bham.ac.uk

[1] Though the cases are imagined, I take them to be realistic, given empirical work on gender and race bias, some of which is discussed below.

http://dx.doi.org/10.1016/j.concog.2014.10.006
1053-8100/

cognitions that have obvious epistemic faults, but also have some significant epistemic benefits that could not be otherwise obtained. In section four, I argue that confabulatory explanations of actions guided by implicit bias have the potential to deliver epistemic benefits, insofar as they maximise the acquisition of true beliefs in the long run by filling an explanatory gap, and they help the agent maintain consistency among her cognitions. I also argue that, at the time of the confabulatory explanation, no alternative explanations are available—in a sense to be explained—to the subject. I conclude that epistemic evaluation of confabulatory explanations should be indexed to context, taking into account the (un)availability of alternatives and potential epistemic benefits. This allows us to resist a kind of trade-off view about the epistemic status of confabulatory explanations: the view that pragmatic benefits come at the expense of epistemic ones. A closer focus on the potential epistemic benefits of these cognitions, as well as the context in which they occur, can result in a more careful epistemic evaluation of them.

1. Explanations of actions driven by implicit bias: two cases

In this section I describe two imaginary cases of confabulatory explanations that could occur in the non-clinical population. I will assume Jules Holroyd's definition of implicit bias in the discussion which follows, according to which: '[an] individual harbors an implicit bias against some stigmatized group (G), when she has automatic cognitive or affective associations between (her concept of) G and some negative property (P) or stereotypic trait (T), which are accessible and can be operative in influencing judgment and behavior without the conscious awareness of the agent' (Holroyd, 2012: 275). Implicit biases then can be understood as 'largely unconscious tendencies to automatically associate concepts with one another', such tendencies can result in judging 'members of stigmatized groups more negatively' (Saul, 2012b: 244). Care is needed when using the term 'unconscious' with respect to implicit bias. Gawronski, Hofmann, and Wilbur (2006) distinguish three types of awareness: *source, content,* and *impact* awareness. If a subject has source awareness of some attitude of hers, she has awareness of the origin of that attitude. If a subject has content awareness of some attitude of hers, she has awareness of the attitude itself. Finally, if a subject has impact awareness of some attitude of hers, she has awareness of the influence of that attitude on other psychological processes (Gawronski et al., 2006: 486). Gawronski and colleagues' review of empirical evidence suggests that implicit attitudes only differ from explicit attitudes with respect to *impact awareness.* They conclude that

> the term "unconscious" is adequate for indirectly assessed attitudes only with regard to one particular aspect: impact awareness. However, the term "unconscious" is inadequate when it is assumed to imply a lack of source awareness or content awareness.
>
> [Gawronski et al. (2006: 496)]

Note that though the evidence suggests that source and content awareness of implicit attitudes is *possible*, that is not to say that subjects always have such awareness in ordinary settings. We can be cautious here and take a lesson from the nearby literature: in her discussion of moderating automatic stereotypes, Irene Blair claims that though 'the evidence is compelling with regard to the possibility of moderating automatic stereotypes, the likelihood of such moderation in everyday social encounters is not yet known' (Blair, 2002: 249). Similarly then, though the work Gawronski and colleagues reviewed showed evidence that content and source awareness is *possible*, that is not to say that 'unconscious' used in this way is always inappropriate.

Empirical work has shown that implicit biases are held by most people, even those who avow egalitarian positions, or are members of the targeted group. Results from Implicit Association Tests (IATs) support this. IATs work by measuring the speed at which subjects pair two categories of object with, for example, pleasant and unpleasant stimuli (e.g. the words 'wonderful' and 'awful') or stereotypical and unstereotypical stimuli. The idea behind the tests is that we can discover which categories a subject associates with one another. This is done by measuring the categorisation performance of combinations of categories (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009: 347). In a review of over 2.5 million IAT results across seventeen topics, Brian Nosek and colleagues report that '[i]mplicit and explicit comparative preferences and stereotypes were widespread across gender, ethnicity, age, political orientation, and region' (Nosek et al., 2007: 40). In the rest of the paper, I will follow Holroyd in understanding implicit biases as *operative* when they 'produce a distorting influence on judgement and hence behaviour informed by that judgement' (Holroyd, 2015). The cases I introduce in the next two sections are ones in which implicit biases are operative in this sense.

1.1. Implicit gender bias and the case of Roger

In a study on the influence of the gender of an applicant on reviewing CVs, both male and female participants were more likely to vote in favour of hiring a male applicant than a female applicant when the CVs presented were otherwise identical (Steinpreis, Anders, & Ritzke, 1999: 509). The authors of the study suggest that the results 'indicate a gender bias for both men and women in preference for male job applicants' (Steinpreis et al., 1999: 510). Both male and female participants were 'significantly more likely to hire a potential male colleague than an equally qualified potential female

colleague', and were 'more likely to positively evaluate the research, teaching, and service contributions of a male job applicant than a female job applicant with an identical record' (Steinpreis et al., 1999: 522). Corinne A. Moss-Racusin and colleagues found that when assessing application materials of a student applying for a Laboratory Manager position, '[f]aculty participants rated the male applicant as significantly more competent and hireable than the (identical) female applicant', as well as 'select[ing] a higher starting salary' (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012: 16,474). Monica Biernat and Diane Kobrynowicz predicted that 'it would be more difficult for women than men [...] to document their ability in a competence-related domain' (Biernat & Kobrynowicz, 1997: 554). They ran two studies which confirmed this prediction showing that women were required by participants to 'jump through more hoops' in order to show that they were able to fill a position (Biernat & Kobrynowicz, 1997: 554). These studies suggest that '[w]ithout any intention of bias, once we have categorised someone as male or female, activated gender stereotypes can colour our perception' (Fine, 2010: 56).

Having pointed to empirical work showing the presence and effects of gender bias, here is the case of Roger:

> *Roger is on a hiring panel deciding from a stack of CVs which candidates to invite to interview. Roger thinks of himself as an egalitarian, and not as somebody who is sexist. The CVs are not anonymous with respect to gender. Roger chooses not to invite any female applicants to interview. Katie is one of the female candidates who Roger chooses not to invite to interview. Katie's CV is of equal or better quality than at least some of her male competitors who did get invited to interview, and had Katie's CV been headed with a male name, Katie would have been invited to interview.*

The case might be made more plausible if the job were a traditionally male one, say, an 'executive chief of staff' position (Biernat & Fuegen, 2001: 711). As Madeline Heilman points out, 'research has repeatedly demonstrated sex bias in employee selection processes [...] with male applicants generally recommended for hire and seen as more likely to succeed than female applicants with the identical credentials when jobs are male in sex-type' (Heilman, 2001: 660). Let us say that Roger's decision not to invite Katie to interview was guided by his implicit bias against women. I will not worry too much about how exactly to cash out the *guided by* locution here, the thought is only that Roger's implicit bias was efficacious in his decision not to invite Katie to interview, and the appropriate counterfactuals are true (*if* Roger did not have an implicit bias against women, Katie *would* have been invited to interview, and *if* Katie's CV was headed with a typically male name, Katie *would* have been invited to interview).

Let us ask Roger to explain his decision. He claims that Katie's CV indicates that she is not good enough for the job, and so she ought not to be interviewed. Later I will argue that Roger's explanation of his decision not to invite Katie to interview is a confabulatory one (Section 2.6).

1.2. Implicit race bias and the case of Sylvia

Empirical work has shown that implicit race bias is exhibited by many people. Such biases may affect not only our beliefs, but also the way we perceive the world. A study by Keith B. Payne (2001) showed that participants were able to more quickly identify guns (as opposed to non-gun tools) when they were primed with a Black face, as compared to when they were primed with a White face. In this experimental set up, the 'presence of Black faces facilitated the identification of guns relative to the presence of White faces' (Payne, 2001: 185). When a time condition constraint was introduced into an otherwise identical experiment, participants more often mistakenly identified tools as guns when they had been primed with a Black face as compared to when they had been primed with a White face. Payne identifies as the 'critical finding' that 'priming participants with a Black rather than a White face was sufficient to make them call a harmless item a gun' (Payne, 2001: 188). In a similar experiment in which participants were instructed to shoot at subjects holding a gun, and not shoot at subjects not holding a gun, participants were found to shoot an armed subject quicker if that subject was African American than if he was White, and the decision to not shoot an unarmed subject was made more quickly when the target was White and not African American (Correll, Park, Judd, & Wittenbrink, 2002: 1317). Finally, behaviour which is ambiguously aggressive committed by a black person is more likely to be perceived as hostile than the same behaviour committed by a white person (see for example Duncan, 1976; Sagar & Schofield, 1980). Implicit bias then is 'getting to us even before we get to the point of reflecting upon the world—it affects our very perceptions of that world' (Saul, 2012b: 246). In summarising a body of work on this Tamar Gendler claims that '[h]undreds of studies conducted in dozens of laboratories over nearly two decades have shown that—for most twentieth and twenty-first-century American participants in most circumstances' the White-or-positive and Black-or-negative pairing 'is more natural' than the White-or-negative and Black-or-positive pairing. Gendler claims that this suggests 'that the former categories are, for most participants, more readily constructed or more easily accessed, and hence experienced as more "natural" than the latter' (Gendler, 2011: 52). Here is the case of Sylvia:

> *Sylvia is walking down the road on her way to work. Sylvia thinks of herself as an egalitarian, and not as somebody who is racist. She sees a black man walking towards her. Sylvia crosses the road. The man is not acting in a threatening manner. Had the man not been black, Sylvia would not have crossed the road.*

Let us say that Sylvia's decision to cross the road was guided by her implicit bias against black people. Again, what I mean by the *guided by* locution here is just that Sylvia's implicit bias against black people was efficacious in the production of her behaviour, and that the appropriate counterfactuals hold (*if* Sylvia did not have an implicit bias against black people, she *would not* have crossed the road, *had* the man not been black, she *would not* have crossed the road).[2]

Let us ask Sylvia to explain her behaviour. She claims that the man was behaving in a threatening way, and so she crossed the road out of fear for her safety. I have built into the case that the black man is not behaving in a threatening way, and let us suppose that if one did not hold an implicit bias towards black people, no threat would be felt. In the next section I will argue that Sylvia's explanation of why she crossed the road is a confabulatory one (Section 2.6).

Next I will look at the explanations offered by Roger and Sylvia in the light of five features of confabulatory explanations, to support the claim that at least sometimes, explanations of decisions or actions guided by implicit bias are confabulatory.

2. Common features of non-clinical confabulation

Here I do not aim to solve the thorny issue regarding how best to define confabulation, rather, I will just highlight what the two cases described in the previous section have in common, focusing in particular on those features that are relevant to their epistemic status. As we will see, these features appear in popular accounts of confabulation. I will consider five common features of confabulatory explanations, they (1) are false or ill-grounded; (2) are offered as the answer to a question; (3) have a motivational component; (4) fill a gap, and (5) are reported without any intention to deceive. These are not necessary and sufficient conditions, rather, they have been thought to be common features of confabulatory explanations, and they are features which characterise the cases of Roger and Sylvia.

2.1. False or ill-grounded

Confabulatory explanations are epistemically faulty. Generally speaking, they are *false* explanations, their being so has been identified as a key or defining feature of them (see for instance Berrios, 2000: 348; McKay & Kinsbourne, 2010: 289). However, someone might confabulate and hit on something true (see for example William Hirstein's discussion of subjects with Korsakoff's syndrome—a form of amnesia caused by alcohol abuse or severe malnutrition—who happen to confabulate correctly (2009: 3), or Ryan McKay and Marcel Kinsbourne's example of a subject who lacks 'access to his biographical information, yet by chance may confabulate the correct answer when asked his age' (2010: 289)). We cannot then, rule out by definition the possibility of true confabulatory explanations. But even when confabulatory explanations are not false, they are epistemically poor in other respects. A key epistemic feature of confabulations then is that they are *ill-grounded* or *poorly supported by evidence* (Hirstein, 2005: 33–4).

In confabulation, the explanation offered for a decision or action does not reflect what caused that decision or action and also poorly matches the details of the situation at hand. For example, the participants in Johnathan Haidt's experiment were presented with a scenario in which two siblings engage in incest. Asked how they felt about the scenario, most of the participants claimed that it was wrong, and justified that attitude on the basis of the risks of inbreeding, even when the case of incest they were presented with explicitly excluded the possibility of reproduction after sexual intercourse (Haidt, 2001).

Explanations of actions guided by implicit bias look to share the same features as the classical cases of non-clinical confabulation reported by Haidt. They are false explanations, since they do not map on to what was efficacious in guiding the subjects' decisions or actions. They are ill-grounded, insofar as they misrepresent important features of the situations. Roger chooses not to invite Katie to interview, when her CV is of equal or better quality than at least some of her male competitors. Sylvia crosses the road upon seeing a black man, when the black man is not acting in a threatening way. Roger offers an explanation of his decision by claiming that Katie's CV was not as strong as the CVs of the (male) candidates who were invited to interview. Sylvia explains her action by claiming that the black man was behaving in a threatening way. The reasons given in these explanations were not the reasons which were actually in play with respect to Roger's decision and Sylvia's action, and they are not supported by evidence, and therefore the explanations are both false and ill-grounded.

2.2. Provoked

In the psychological literature there is a recurrent distinction between two types of confabulation. A confabulation is *spontaneous* when it is not elicited by questioning, and it is *provoked* when it is offered as a response: the person is asked to offer an explanation for something, and she provides a confabulation (Kopelman, 1999: 197–8; Hirstein, 2005: 20).

[2] In some cases it is difficult to specify the relevant counterfactual, for example, those cases in which the effect of bias is to hasten or exaggerate a response, rather than to make a difference to the behavioural output. For example, the participants in Payne's studies who correctly identified a gun, but were quicker to shoot if the target was black. In such cases the counterfactuals cannot be cashed out in terms of the behaviour (shooting), since even if the targets were white, the subjects may well have still shot (since the white targets would also have been holding guns). I think this is unproblematic in the cases of Roger and Sylvia, since the behaviour there is clear cut (and not merely hastened or exaggerated). I do not commit to it being necessary that appropriate counterfactuals are true of *all* behaviour guided by implicit biases (though perhaps this is right providing we are prepared to offer very fine-grained counterfactuals in some cases). Certainly in the cases of Roger and Sylvia though, some interesting counterfactuals hold of their behaviour which is guided by bias. I am grateful to Jules Holroyd for raising this worry.

Provoked confabulations can occur in people 'who are fully in possession of most of their cognitive faculties, and able to respond correctly to all sorts of requests and questions' (Hirstein, 2005: 21).

In the two cases of confabulation I considered in the previous section and in many other instances of non-clinical confabulation, the explanation is likely to be generated as a response to a specific question, such as 'Why did you not invite this candidate to interview?' or 'Why did you cross the road?' The second feature of confabulation then, which is shared by the cases of Roger and Sylvia, is that they can be provoked as responses to questions.

2.3. Motivated

A third feature common to many instances of confabulation is that they involve a motivational element, which means that they are goal-directed states or processes (Bayne & Fernandez, 2009). This is something 'many accounts of confabulation' have recognised (Bortolotti & Cox, 2009: 954, see also Zangwill, 1953: 700). This motivational element might be a causal factor for the explanation's being provided *at all* (for example, to cover a gap in memory, Bonhoeffer, 1901, cited in McKay & Kinsbourne, 2010: 291), or as a factor with respect to the *content* an explanation has (Fotopoulou, Conway, & Solms, 2007; Fotopoulou et al., 2008; Metcalf, Langdon, & Coltheart, 2010). So disagreement with respect to the role of motivational factors in confabulation concerns whether they play a role in the very existence of a confabulation, or whether they play a role additionally in the content of a confabulation.

Confabulatory explanations are then, at least in part, *motivated* explanations, that is, some pro-attitude plays a role in either the very formation of a confabulation, or more specifically, the content of a confabulation. We can distinguish between a motivation to offer an explanation as opposed to no explanation (type-a), and a motivation to offer an explanation with a specific content as opposed to an explanation with a different content (type-b).

In the non-clinical context of confabulatory explanations of decisions or actions guided by implicit bias, confabulatory explanations may be motivated in both senses. In some cases, they allow the person to avoid the appearance of ignorance or incompetence. In response to a question to which a person does not know or cannot know the answer, she may confabulate when admitting that she does not know the answer would be costly. This typically occurs when 'the provoking question touches on something people are normally expected to know' (Hirstein, 2005: 30), (we might think that someone on a hiring panel ought to be able to give reasons for rejecting candidates, and someone crossing a road ought to be able to give reasons for so doing). Thus, confabulatory explanations may protect a person from claiming not to know why they decided or acted as they did. Equally, there are cases in which the motivational component plays a role in the very content of the explanation. Perhaps the most obvious motivational component guiding the content of a confabulation is the desire a subject might have to maintain coherent beliefs about the self or a good self-concept. It looks like both kinds of causal work are being done by confabulatory explanations offered in the cases of Roger and Sylvia. In the first place, the explanation might be motivated by a desire not to be dumbfounded, or the desire not to fail to offer an explanation at all for one's decision or action. In the second place, the content of the explanation might be motivated by one's desire not to appear sexist or racist.

2.4. Filling a gap

Related to the last point, and especially with regards to type-a motivation, confabulatory explanations have been thought to, in some sense, *fill a gap.* According to the classic conception of clinical confabulation as a consequence of a memory impairment—that best applies to the phenomenon when it occurs in the context of amnesia and dementia—confabulations are 'stories produced to cover gaps in memory' (Hirstein, 2005: 32). But confabulation plays a gap-filling role in more cases than those in which there is a memory deficit (I do not think that Roger and Sylvia offer the explanations they do because they fail to remember why they acted as they did[3]). The gap-filling claim needs to be construed more broadly and not couched only with respect to a gap in *memory.* We might instead think of confabulatory explanations as filling gaps 'at a certain level in the cognitive system', insofar as they help produce 'complete, coherent representations of the world' (Hirstein, 2005: 30).

In the cases I have given of explanations of actions guided by implicit bias, the person may not be aware of having a bias, or at the very least, may not be aware of that bias's influence, (recall that this is what Gawronski and colleagues term *impact awareness* (2006: 486)). This is the gap that might be filled by the confabulatory explanation. Arguably, explanations of decisions or actions guided by implicit bias fill a gap which could not be otherwise filled—I will discuss the idea of alternative explanations being in some sense *unavailable* to subjects who confabulate later (Section 4.2).

2.5. No intention to deceive

My talk of a motivational component and of gap-filling may be suggestive of an intention to deceive (even if it is an intention to deceive *oneself*) when subjects offer confabulatory explanations. There is an interesting literature on the potential overlap between confabulation and self-deception which I cannot address here (see, for example Bayne & Fernandez,

[3] There is of course a sense in which Roger and Sylvia fail to remember—the sense in which one fails to remember something one was never aware of. This is not the sense of failure to remember I am interested in here. Though Roger and Sylvia do fail to remember in the sense that they are *unable* to remember, this is because the un-remembered explanation of their actions is not one which they were ever aware of (rather than them being aware of the explanation at some past time, and failing to recall at the time at which they confabulate). I am grateful to Jon Robson for encouraging me to make this clearer.

2009; Hirstein, 2000, 2005; Ramachandran, 1996). For my purposes it suffices to say that the 'orthodox position' is that people who confabulate should not be understood as *lying* (Hirstein, 2005: 28), and this follows from there being no intention to deceive in confabulation.[4] This is made very clear in the epistemic concept of confabulation, according to which a confabulation is 'a certain type of epistemically ill-grounded claim that the subject does not know is ill-grounded' (Hirstein, 2005: 33). According to this conception, when a person confabulates she offers an ill-grounded claim *and* she is not aware of that claim *as* being ill-grounded, and she does not believe contrary to it. If a necessary condition on intending to deceive another is that the deceiver believes contrary to what she avows, then she who confabulates does not so intend, because she does not recognise her confabulation as epistemically ill-grounded.

People offering confabulatory explanations of decisions or behaviour guided by implicit bias have no intention to deceive: subjects can lack impact awareness of the operations of implicit biases, and so it would be strange to understand the subjects in my examples as seeking to deceive in order to cover up their implicit sexism or racism which is operative when they act. Given that their attitudes are implicit ones and are not recognised by the subjects as attitudes they have (or at least, not recognised as attitudes which are influencing their decisions or behaviour), subjects should not be construed as intentionally seeking to cover up those attitudes (see Holroyd, 2015, for more discussion on this point).

2.6. Do Roger and Sylvia confabulate?

Are the explanations given by Roger and Sylvia genuine cases of confabulation? I have outlined five features common to confabulation, these are their being false or ill-grounded, provoked in response to questioning, motivated, filling a gap, and there being no intention to deceive on the part of the confabulator. In my outline of each of these features, I have suggested that they are ones which characterise the cases of Roger and Sylvia.

Care needs to be taken though with respect to how the cases of Roger and Sylvia are described, since under a certain kind of description, Roger and Sylvia are not confabulating. In the cases as I have described them Roger genuinely (and falsely) believes that Katie's CV is of inferior quality, and Sylvia genuinely (and falsely) believes that the black man is behaving in a threatening way. If this is the interpretation of the circumstances—that the subjects have false beliefs and base their decisions and actions upon those beliefs—we would not be in the presence of confabulatory explanations. Our subjects would be explaining their behaviour by appealing to a belief of theirs for which the evidence was poor. Under the hypothesis I am currently discussing, Roger believes that Katie has a worse CV than her male competitors and he offers a true explanation of his action when he explains why he did not choose to invite her to interview. His belief about the quality of the CV is not supported by the evidence, but instead is guided by an implicit bias, the belief is false but the explanation of the choice is not confabulatory. The explanation is one which is both true and well-grounded. This is not how I want to understand these cases. Rather, let us ask Roger why he has the belief that Katie's CV is of inferior quality. Now Roger claims that he has the belief because the CV *is* inferior. Now we *are* in the presence of confabulatory explanations since the reason Roger has the belief that the CV is inferior is not because it is, it is rather because of some implicit bias he has. My cases then ought to be understood as ones in which the subject is asked to explain why they acted in the way that they did, and they give an explanation in which they cite what they take to be a fact about the world which both informs the belief and explains the decision or action. When Roger says that he did not invite Katie to interview because her CV was inferior he confabulates, because he cites something which was not efficacious in the making of his decision. This is not to say that it is sufficient for confabulation that one is wrong about the cause of one's belief, and it is for this reason only that Roger's explanation is a confabulatory one. As we have seen, Roger's explanation exhibits many other features typical of confabulatory explanations.

Note also the way the cases were described: these are not cases of *explicit* sexism or racism, rather, they are cases of subjects who explicitly take egalitarian positions, but have *implicit* biases against certain groups, which guide their decisions or actions. It is possible (indeed common) for a subject to have no explicit sexist or racist beliefs and yet still have implicit biases towards these groups. According to the most common reading of the IAT results, subjects have implicit biases, even though such biases are not always reflected in explicit judgements.

3. Epistemic innocence[5]

I am interested in the epistemic status of confabulatory explanations of decisions or actions guided by implicit bias, and whether they have the potential for *epistemic innocence*. I will understand the notion of epistemic innocence in the following way: an epistemically faulty cognition is epistemically innocent if, at a given time, it endows some significant epistemic benefit (*Epistemic Benefit*) onto the subject, which could not be otherwise had, because alternative, less epistemically faulty cog-

[4] Recently some philosophers have argued that it is not a necessary condition on lying that one intends to deceive, and if this is right, the fact that people who confabulate do not intend to deceive does not—at least not by itself—support the claim that confabulators are not liars. However, alternative accounts of lying do seem to agree on another condition on lying which people who confabulate also fail to meet, that is, asserting something that one does not believe, or more strongly, something which one believes to be false (see for example, Carson, 2010; Fallis, 2009; Sorenson, 2007. See also Lackey (2013) for an overview of the recent literature).

[5] Thank you to Lisa Bortolotti, with whom I developed the notion of epistemic innocence as I am understanding it here (see Bortolotti, 2015, for an explanation of why we use the term 'innocence' in this way).

nitions are in some sense *unavailable* to her at that time (*No Alternatives)*. I will further elucidate the notion of epistemic innocence in the rest of this section.

3.1. Which benefits?

A cognition being epistemically innocent does not imply that that cognition is free from epistemic faults, perhaps very few cognitions are innocent in that sense. The claim is only that such cognitions can confer epistemic benefits which are otherwise unavailable, in some contexts.

If confabulatory explanations of the kind I have in mind here have epistemic benefits, might we be better off to call them epistemically *good* rather than epistemically *innocent*? No, since saying this would be to ignore or deny the obvious epistemic faults of confabulation. My aim here is to highlight the potential epistemic benefits of some confabulatory explanations. One way of doing this without polarising the debate is to claim some sort of inbetween status for confabulations. An ideal agent would not need to confabulate explanations, but human agents have significant limitations that lead to confabulatory explanations. Now, this is not always or not entirely a bad thing. Confabulatory explanations can sometimes play a positive epistemic role.

3.2. Which unavailability?

What I mean by *unavailable* in the No Alternatives condition on epistemic innocence will differ depending on the kind of cognition under investigation. Here is a first pass at three ways we might think about the notion of unavailability: alternative cognitions might be strictly unavailable, motivationally unavailable, or explanatorily unavailable.

An alternative explanation is *strictly unavailable* if it is based on information that is opaque to introspection, or otherwise irretrievable. We might understand this sense of unavailability in terms of alternative cognitions being *inaccessible* to the subject. For example, consider the case of a subject with dementia who suffers from severe memory impairment. She claims to remember going to the beach with her parents that morning, but the trip she recalls occurred when she was a teenager, sixty years ago. A memory of the trip which included the correct time at which it took place, or information which would suggest to the subject that she had made a mistake with respect to the time of the trip, is *inaccessible* to her, and so strictly unavailable, due to the severe memory impairment she suffers as a result of her dementia.

An alternative explanation is *motivationally unavailable* if it is inhibited or not accessed due to motivational factors. It is generally agreed upon in the literature that self-deception includes a motivational element, which makes it a good case to refer to in explicating the notion of motivational unavailability.[6] This motivational element might make less epistemically faulty cognitions unavailable to the subject. Take the case of the cuckholded husband who self-deceptively believes that his wife is faithful. Evidence that she is unfaithful may be available to him (insofar as it is perceptually available—he sees that his wife returns home late, dishevelled, and uninterested in him), but an alternative cognition, such as the belief that his wife is having an affair, is motivationally unavailable, due to the husband's very strong motivation for it to be the case that his wife is faithful (wishful self-deception) or for it to be the case that he *believes* that his wife is faithful (willful self-deception) (see Van Leeuwen, 2007: 331–2, for further elucidation of the distinction between wishful and wilful self-deception).

An alternative explanation might be unavailable in a weaker sense, such that it is strictly speaking available for consideration by the subject, but it is not regarded as a genuine contender. It is *explanatorily unavailable* to the subject insofar as it is dismissed due to its apparent implausibility. For example, a subject may come to have a cognition which explains some experience she has. If alternative cognitions which might also be candidate explanations for her experiences are such that they strike her as seriously implausible or explanatorily inadequate, these alternative cognitions are *explanatorily unavailable*. Suppose there are bite marks in my cheese, I hear scratching at night, and my cat is agitated. I come to the conclusion that I have mice in my house. An alternative explanation might be that a cheese-eating, cat-irritating, noisy fairy is infiltrating my home at night. This explanation is not available to me in the sense I have in mind here due to the incredulity I would feel towards it. It is either not considered by me, or it is such that I rule it out on grounds of implausibility or poor explanatory power, relative to the preferred and adopted cognition.

4. Epistemic innocence and confabulatory explanations

In order for confabulatory explanations of decisions or actions guided by implicit bias to be epistemically innocent in the way described above, it would need to be the case that they are epistemically beneficial as per the *Epistemic Benefit* condition,

[6] One might have something like the following worry about my notion of *motivational unavailability.* Perhaps what I have in some cases, in self-deception for example, is motivational factors affecting the threshold for belief that *p, not* making alternative beliefs in any sense *unavailable,* so the idea of motivational availability as ascribed to cases of self-deception is not quite right. It is not that some motivational state makes other cognitions *unavailable*, rather it gives more epistemically worthy cognitions less weight, though they are still *available*. If one is inclined to understand self-deception in these terms, then one can simply re-describe motivational unavailability in terms of the priorities to search for alternative hypotheses one is inclined to have. If we are inclined to view motivational factors in this light, motivational unavailability collapses into explanatory unavailability, so that it is not the case that alternative cognitions are *unavailable,* strictly speaking. Rather, the motivational factors involved are such that those alternatives weigh much less heavily with the subject than they epistemically ought to. I am grateful to Paul Noordhof for raising this worry.

and that less epistemically faulty alternative explanations delivering the same epistemic benefit are in some sense unavailable to the subject at the time, as per the *No Alternatives* condition. In this section, I will argue that at least in some cases, confabulatory explanations of this sort meet both conditions on epistemic innocence.

4.1. Epistemic benefit and confabulatory explanations

As in my earlier characterisation, confabulatory explanations are epistemically poor insofar as they are false, or at the very least ill-grounded. Here I want to focus on whether confabulatory explanations have any epistemic benefits *notwithstanding their epistemic faults*. I claim that there are two potential epistemic benefits: in the context of epistemically imperfect agents, confabulatory explanations of decisions or actions guided by implicit bias may maximise the acquisition of true beliefs in the long run by filling an explanatory gap, and they may help the epistemic agent maintain consistency among her cognitions. Sometimes confabulations will be epistemically innocent and overall good, for example, viewed from a consequentialist framework this might be when the benefits outweigh the costs. In other cases, they will be epistemically innocent, and epistemically bad, from the same framework, as when such benefits are counteracted by the epistemic costs of confabulation. Importantly, epistemic innocence does not stand or fall with epistemic goodness.

4.1.1. Maximising the future acquisition of true beliefs by filling explanatory gaps

Here I will argue that there are epistemic benefits in providing a confabulatory explanation to the question 'Why did you do this?' This is because filling an explanatory gap allows people to consider their reasons for their actions which may otherwise remain unexplored and unchallenged. Earlier I suggested that one of the characteristics of confabulatory explanations was their filling a gap, where this was to be understood not just in terms of memory gaps, but rather, more broadly. The performance of the function of gap filling may be pragmatically beneficial in different ways depending on the case. In the non-clinical case, explanations of decisions or actions guided by implicit bias may fill a gap which gives the illusion of competence, and prevents people from having to claim that they do not know or cannot remember why they formed an attitude, made a choice, or performed an action (see Dalla Barba, 1993, cited in Van Damme & d'Ydewalle, 2010: 221).

McKay and Kinsbourne identify two types of motivational account which apply to confabulations. According to the first, subjects who confabulate offer the explanations they do 'in order to conceal embarrassing gaps in their memories' (McKay & Kinsbourne, 2010: 291). McKay and Kinsbourne reject this account as one having 'very little evidence' in support of it, and I have already said that confabulations are doing more than filling gaps in memory. We might understand this kind of account more broadly though, as one which claims that people who confabulate do so in order to conceal gaps more generally (for example, explanatory gaps regarding why they decided or acted in a certain way). If this kind of account were correct, at least for confabulatory explanations in the non-clinical population, such explanations could be seen as conferring a pragmatic benefit with respect to the prevention of embarrassment which may have indirect positive epistemic consequences (perhaps my not suffering the discomfort of embarrassment might mean I am more willing and able to investigate my environment and participate in the exchange of information).

According to the second type of motivational account, confabulations are 'purposive constructions that function to embellish the situation of the patient', meaning that confabulatory explanations are *compensatory* in virtue of their content (not merely their very existence) (McKay & Kinsbourne, 2010: 291). If the second type of account were right, there would be additional pragmatic benefits, due to the confabulation enhancing the concept of the self or the situation of the person, and potentially enhancing self-confidence and wellbeing. This matches with my description of the cases Roger and Sylvia. But is there room for *epistemic*, as well as pragmatic, benefits?

My claim here is that by having an explanation and endorsing a position, one can receive feedback and the problematic things that are believed (*I did not invite Katie to interview because her CV was of poorer quality* or *I crossed the road because the black man was behaving in a threatening way*) become available to introspective reflection, and open to feedback and revision. Jeanette Kennett and Cordelia Fine argue that when we become aware of our biases and are committed to not being prejudiced, then we can counteract or compensate for our biases and achieve better consistency between explicit attitudes and behaviour:

> research demonstrates that when people become aware that they have a tendency to make certain types of judgments in a biased way (for example, due to the activation of negative stereotypes about a racial group) then, if they are motivated to be unprejudiced, they will effortfully over-ride their intuitively-based judgments, so long as they have the cognitive resources to do so.
>
> [Kennett and Fine (2009: 89)]

If becoming aware of our biases and being motivated to not be prejudiced can lead to trying to overcome the judgements we make on the basis of bias, confabulatory explanations may be indirectly epistemically beneficial, insofar as they might contribute to our becoming aware of our biases. We might think that a much better way to become aware of our biases would be to fail to offer an explanation at all. In Roger's case for example, if he claimed not to know why he did not invite

Katie to interview, this might be a better way of encouraging him to reflect on what was guiding that decision. However, I will argue later (Section 4.2) that this dumbfounding response is *unavailable* to Roger, and so the confabulatory explanation here confers a benefit which is not otherwise obtainable.

Related to this, research has shown that individuals who avow low prejudice but recognise that they are prone to behaviour which is inconsistent with this—so-called 'low-prejudice discrepancy-prone' subjects—are more likely to feel guilt when expressing biased behaviour (as measured by the IAT for race), and more likely to interpret such behaviour as being related to race factors (Monteith, Voils, & Ashburn-Nardo, 2001: 411). Speculatively, if subjects become aware that they are discrepancy-prone, the guilt felt when expressing biased behaviour, and their interpreting it in this way, might be instrumental in seeking to achieve better consistency between their explicit attitudes and their behaviour. There are reasons to think such awareness is possible. For example, in a study looking at the relationship between self-report scores on the Modern Racism Scale (MRS) and implicit racial attitudes, Jason Nier found that:

> when participants believed their 'true attitudes' were being accurately assessed, there was a significant relationship between an implicit measure of racial attitudes (the IAT) and an explicit measure of racial attitudes (the MRS). When participants did not believe that their self-reported explicit attitudes could be accurately corroborated with an implicit measure, there was no association between implicit and explicit attitudes.
>
> [Nier (2001: 48–9)]

One way to become aware of our biases is to observe our own behaviour (see Holroyd, 2015) compare them with our explanation of our actions, and identify discrepancies. Giving reasons (even confabulatory ones) contributes to becoming aware of one's own attitudes (and conflicts in attitudes), priorities and values and offers people the opportunity to construct a coherent narrative of themselves where behaviour aligns with explicit commitments as opposed to implicit biases (Bortolotti, 2009). By reporting a false explanation, we make some (confabulatory) explanation for our decisions or actions available, and we may initiate a process by which we acquire a new true belief, whereas if we did not have any explanation to offer we may have been stuck with an implicit bias that is not detected and does not get challenged.

The flip side to this benefit is that offering a confabulatory explanation might serve to mask the discriminatory nature of the behaviour, which might otherwise be observed if the explanation were an accurate one or one indicating dumbfounding. I will argue later though that alternative explanations such as these are unavailable to the person confabulating, given certain other conditions (Section 4.2). I do not deny the epistemic costs of the confabulatory explanation—that it might mask the discriminatory nature of the behaviour being one such cost—the idea is only that there may also be some epistemic benefits, including the confabulatory explanation initiating a process by which the subject acquires a new true belief.

So confabulatory explanations are epistemically beneficial insofar as in offering them as explanations they become open to attack. This might then start a thinking process, where I might end up with a less epistemically faulty explanation. So the confabulation is epistemically good insofar as it acts as an enabling condition for further reflection. Discussing and justifying one's decisions or actions, specifically, reflecting on the potential role played by stereotypes can reduce the effects of such stereotypes (Saul, 2012a: 259). This does not require the subject to have any kind of conscious control over her implicit biases, the idea is rather that when she engages in this kind of discussion, it 'may help to flag up at least some cases in which there is no defensible reason for a judgment. And some of these may be cases of bias' (Saul, 2012a: 259). This can be epistemically beneficial insofar as if one discovers that the decision or action is not carried out for a defensible reason this might initiate a process whereby the subject comes closer to a true explanation of their decision or action. The subject may come to have impact awareness of their bias, and seek to reduce its influence which might have positive epistemic consequences.

One worry is that the benefit I have outlined is one had by a lot of things, and that now my just saying anything false is epistemically beneficial insofar as it might prompt someone to suggest I think again, which might in turn bring about my doing so, and then I might come to have a more epistemically worthy belief. There are two things to say in response to this kind of worry. Firstly, this is an epistemic benefit which is had in a context in which alternatives are not available, whereas not all cases of avowed false belief are going to be cases where less epistemically faulty cognitions are unavailable. So though avowing false beliefs might be epistemically beneficial, the context in which this benefit comes is important for judgements of epistemic innocence. Secondly, it is not on this benefit alone that I suggest that confabulatory explanations of decisions or actions guided by implicit bias are epistemically beneficial.

4.1.2. Maintaining consistency

Here I will argue that there are epistemic benefits in providing a confabulatory explanation in response to a question such as 'Why did you do this?' because it allows one to maintain a coherent self-concept. The epistemic benefit here comes from the explanation's having a particular content, and so this maintaining of consistency is only had by explanations with certain contents.

Not having unexplained gaps with respect to one's decisions or actions may be instrumental to maintaining a coherent set of beliefs about oneself. In the case of implicit bias, Roger has to make consistent his belief that he is egalitarian, and his belief that he did not invite Katie to interview. Similarly, Sylvia has to make consistent her belief that she is egalitarian, and her belief that she crossed the road when she saw a black man. Given our subjects' lack of impact awareness that an

implicit bias is playing a role here, their confabulatory explanations maintain consistency between their beliefs about the values they are committed to, and their beliefs that they made some decision or performed some action.[7]

In addition to this, the explanation may fill a gap in the sense that it provides a reason for the action that is more desirable than the (true) alternative. In the confabulatory explanations of actions driven by implicit bias, Roger's decision not to invite Katie to interview, or Sylvia's action of crossing the road to avoid a black man, is explained by appeal to something external (the inferior quality of Katie's CV, and the threatening behaviour of the black man, respectively). The subjects make an attribution error but in doing so they maintain their positive self-concept. In the true explanations, the action would be caused by something internal, a bias against women, and a bias against black people, respectively.

The limitations of this potential epistemic benefit lie in the fact that the confabulation achieves consistency at the expense of truth. This may lead to a revision of the self-image such that in confabulation, 'coherence trumps correspondence' (Bortolotti & Cox, 2009: 962).[8] We might wonder then why consistency here should be considered as epistemically beneficial, might it make the acquisition of true beliefs less likely in the long run? If we conceive of consistency being a benefit only when it tends towards truth or when it adds to the overall coherence of belief, perhaps the maintaining of consistency had here is not epistemically beneficial. For those who feel the force of this worry I say this: at the very least the apparent maintaining of consistency might be *indirectly* epistemically beneficial. Not experiencing an inconsistency might have indirect epistemic consequences. Just like the prevention of embarrassment (see Section 4.1.1), perhaps by not suffering the discomfort which an inconsistent set of cognitions might bring, I might be more willing and able to investigate my environment and participate in the exchange of information.

4.2. No alternatives and confabulatory explanations

I have argued that confabulatory explanations can have epistemic benefits. Here I will argue that such benefits are only attainable via the confabulatory explanation, because less epistemically faulty cognitions which deliver the same epistemic benefits are in some sense unavailable to the subject who confabulates.

We have seen that confabulatory explanations of decisions or actions guided by implicit bias can be *sincerely* and confidently avowed, and offered with no intention to deceive on the part of the subject confabulating. Given this, as well as the third-person implausibility of the explanations offered, it looks like subjects who confabulate do not recognise the poor epistemic status of the explanations they give. This claim is supported further by accounts of confabulation which postulate a deficit in the confabulator's ability to evaluate hypotheses. On Hirstein's view, two errors are involved in confabulation. The first is the creation of a false response, and the second is a failure by the subject to 'check, examine it and recognize its falsity' (Hirstein, 2005: 15). The thought is that a non-confabulator might have the first error, the generation of a false response, but they would recognise the falsity or absurdity of it. So the person who confabulates 'should either not have created the false response or, having created it, should have censored or corrected it' (Hirstein, 2005: 15). The mechanisms by which this failure is achieved is suggested by Hirstein to result from the fact that the ability to construct plausible responses and the ability to verify such responses are separate in the brain, and confabulators retain the first ability, but the second has been compromised by brain damage (Hirstein, 2005: 16–7).

My focus has been on confabulation as it occurs in the non-clinical population, specifically, in explanations of actions guided by implicit bias. Hirstein states that young children have been reported to confabulate when reporting on their (false) memories (see for example Ackil & Zaragoza, 1998), and subjects of hypnosis can confabulate (Hirstein, 2005: 16), as well as ordinary people in experimental settings (for example, in choice blindness experiments, see Hall & Johansson, 2008). In these cases, people do not have the kind of brain damage which would make the particulars of Hirstein's account plausible in such a way that it has general application across all cases of confabulation.

The etiology of confabulatory explanations may not be shared across the board; we cannot give an account which cites brain damage when we are explaining why non-clinical subjects offer confabulatory explanations in certain contexts. What we can say in general though is that—for whatever reason—when people confabulate, they do not experience doubt, which results in a failure to check on the plausibility of hypotheses when they are occur. Perhaps this failure is realised in different ways across cases. In a case of explanations of decisions or actions guided by implicit bias, for example, the failure might be caused by some motivational component of the agent's cognitive economy.

The motivational component of explanations of actions guided by implicit bias is straightforward to spell out. In terms of type-a motivation (that is, motivation to offer *an* explanation as opposed to no explanation), subjects are asked to justify their decision or action but because their decision or action was guided by an implicit bias, the true causal explanation of their decision or action is not strictly available to them (unless they are well-read in psychology and they suspect that their

[7] This point relates to the theory of cognitive dissonance, the idea that 'inconsistency between two cognitions creates an aversive state akin to hunger or thirst that gives rise to a motivation to reduce the inconsistency' (Cooper & Carlsmith, 2001: 2112). It is beyond the scope of this paper to discuss dissonance more fully, but particularly relevant here is Elliot Aronson's claim that where dissonance theory 'makes its clearest and neatest prediction' is when we are 'dealing with the self-concept and cognitions about some behavior. If dissonance exists it is because the individual's behaviour is inconsistent with his self-concept' (1968: 23). This characterisation of dissonance sits especially well with my claim that confabulatory explanations maintain consistency between beliefs a subject holds and beliefs about some behaviour performed by the subject. Subjects might be motivated to make these consistent in part because dissonance between them is experienced.

[8] This is a descriptive claim. I am not suggesting that coherence ought to trump correspondence, but only that in some cases of confabulation, this is what seems to occur.

decision or action may be influenced at least to some extent by biases they are not aware of). As we have seen, this lack of impact awareness with respect to implicit attitudes is a distinguishing feature of them (Gawronski et al., 2006: 486). Given that failing to give reasons for one's decision or action may be costly, such subjects have a reason to provide an explanation.

In terms of type-b motivation (that is, motivation to offer an explanation with a particular *content*), our subjects are motivated to think of themselves as persons of egalitarian persuasion, who do not have gender or racial prejudices. Thus, it is easy to see that the explanations they offer for their decisions and actions are influenced by motivational factors. We might think that motivational factors are playing a role in the *content* of the explanation offered by running the appropriate counterfactual: if Roger were *explicitly* sexist or Sylvia were *explicitly* racist, and the subjects were not ashamed by this or did not have the desire to conceal such prejudices, and they decided or acted in the above ways and were asked for explanations of their decisions or actions, we might expect Roger to claim that he did not invite Katie to interview because women make for bad colleagues, and Sylvia to claim that she crossed the road to avoid the black man because black men are dangerous. If a subject's decision or action is guided by an implicit bias pertaining to some group, there is a sense in which explanations citing that bias are unavailable to her, given certain other conditions. Such other conditions may include the subject considering herself a person who does not hold sexist or racist prejudices, or at the very least having the desire to be perceived as such by her peers:

> The belief that one's actions are implicitly biased, and other implied beliefs about one's role in sustaining patterns of discrimination, are clearly beliefs that, for a range of reasons, agents might be motivated not to confront. Conversely, the belief that one's actions are consistent with one's moral ideals (of non-discrimination, of being evidence sensitive and unbiased) is one that agents are motivated to maintain.
>
> [Holroyd (2015)]

This kind of story is supported empirically. In their review article, Gawronski and colleagues found evidence that correlations between self-reported attitudes and indirect attitudes are 'often higher when the impact of motivational factors is controlled' (Gawronski et al., 2006: 489). Confabulatory explanations of actions driven by implicit bias are sometimes such that less epistemically faulty alternative cognitions are motivationally unavailable to the subject insofar as they cite implicit biases against certain groups, biases which she is motivated not to recognise in herself or have others recognise in her.[9]

Thinking in terms of one's attitudes being regulated for truth, we might say that the truth regulation present in the cases when a motivational component is in play is sufficiently different from the truth regulation present in cases where motivational factors are not playing a role, such that in the former case, the content of the explanation offered is distorted away from a true explanation. Once the subject is motivated to think of herself as someone who does not have sexist or racist attitudes, or motivated for other people to not think of her in that way, the regulation for truth in her belief formation with respect to her explanation for her decision or action is weakened. We might think then, that without the motivational component, the explanation offered for decisions and actions guided by implicit biases, would not be confabulatory, since studies have shown that controlling for the motivational components results in a higher correlation of explicit and implicit attitudes (Gawronski et al., 2006: 489).

What cases of confabulation share is that something is going on such that where doubt *should* be cast on the confabulatory story, it is not. My claim is that whatever this component is—whether it has a neurological basis (as in many clinical cases), or a motivational one—it indicates that other, less epistemically faulty cognitions are unavailable to the subject. When an explanation is presented to the subject upon which doubt is not cast, alternative cognitions are unavailable insofar as the confidence which comes with the confabulation, and the motivational factors which are driving both the presence and content of it, close off alternatives such that they are not sought or entertained.

5. Conclusions

In this paper I characterised confabulation in terms of five features which are common to them: their being false, their being offered in response to a question, their involving a motivational component, their filling a gap, and their being produced and avowed without any intention to deceive. After outlining two possible cases of confabulatory explanations of decisions or actions guided by implicit bias, I introduced the notion of *epistemic innocence*. This notion was intended to capture those cognitions which deliver some epistemic benefit which could not have been otherwise had in virtue of alternative cognitions being in some sense unavailable. I then argued that confabulatory explanations of this sort have the potential for epistemic innocence.

[9] This is not to say they could not become aware of these attitudes, indeed, recent work has suggested that people are, or can become aware of implicit attitudes they have. For example, we might suppose that Sylvia, who has implicit attitudes against black people, has an affective response when she perceives a black man, and so crosses the road. Adam Hahn and colleagues found that participants were able to predict their IAT results, even when their explicit attitudes did not coincide with their implicit ones. They hypothesize that the reason participants were accurate in their predictions is because when participants 'were presented with the attitude targets, participants did in fact "feel" their affective reaction and reported on those reactions as their implicit attitudes, even though they might have invalidated those same responses as a basis for their explicit attitudes' (Hahn, Judd, Hirsh, & Blair, 2014: 1387). In the right conditions, and having knowledge of implicit biases, Sylvia might be able to report that she has an implicit bias against black people, let us suppose though that she is not in those conditions, and nor does she have the appropriate knowledge of implicit biases.

So what should we conclude from these considerations about the epistemic status of confabulatory explanations in the non-clinical population? What might my analysis mean for the epistemic evaluation of non-clinical confabulation, specifically confabulatory explanations of decisions and actions guided by implicit bias? As I noted earlier, epistemic innocence does not track epistemic goodness. The benefits I identified though should not be neglected. These were filling an explanatory gap which cannot be otherwise filled, which may lead to the acquisition and retention of true beliefs or knowledge, as well as helping to maintain consistency between a subject's other beliefs (which might be directly or indirectly epistemically beneficial). In some cases, less epistemically faulty cognitions with the same epistemic benefit are in some sense unavailable to the subject, and so the confabulatory explanation delivers some epistemic benefit which is otherwise unobtainable. I conclude then that at least in some cases, confabulatory explanations of decisions or actions guided by implicit bias can be epistemically innocent, and that epistemic evaluation of confabulatory explanations ought to take into account the context in which the cognition occurs. If we do this we are able to provide a richer account of the epistemic status of confabulatory explanations, and we can resist the view that pragmatic benefits come at the expense of epistemic ones. If we focus on the context in which confabulatory explanations occur, and their potential epistemic benefits, we can give a richer epistemic evaluation of them.

Acknowledgments

I acknowledge the support of the Arts and Humanities Research Council (*The Epistemic Innocence of Imperfect Cognitions*, Grant Number: AH/K003615/1). Thank you also to the Mind and Reason research group at the University of York and the audience of the Workshop on Costs and Benefits of Imperfect Cognitions (University of Birmingham, 8th–9th May 2014) for helpful comments on a previous version of this paper. I am particularly grateful to Lisa Bortolotti, Jules Holroyd, Paul Noordhof, Jon Robson, Glen Sullivan-Bissett, Jane Tomlinson, and two referees for extensive helpful comments which greatly improved the paper.

References

Ackil, J. K., & Zaragoza, M. S. (1998). Memorial consequences of forced confabulation: Age differences in susceptibility to false memories. *Developmental Psychology, 34*(6), 1358–1372.
Aronson, E. (1968). Dissonance theory: Progress and problems. In R. P. Ableson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook*. Chicago: Rand McNully.
Bayne, T., & Fernandez, J. (2009). Delusion and self-deception: Mapping the terrain. In T. Bayne & J. Fernandez (Eds.), *Delusion and self-deception: Affective and motivational influences on belief formation* (pp. 1–21). Hove, UK: Psychology Press.
Berrios, G. E. (2000). Confabulations. In G. E. Berrios & J. Hodges (Eds.), *Memory disorders in psychiatric practice* (pp. 348–368). Cambridge University Press.
Biernat, M., & Fuegen, K. (2001). Shifting standards and the evaluation of competence: Complexity in gender-based judgment and decision making. *Journal of Social Issues, 57*(4), 707–724.
Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology, 72*(3), 544–557.
Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*(3), 242–261.
Bonhoeffer, K. (1901). *Die akuten Geisteskrankheiten der Gewohnheitstrinker: Eine Klinische Studie*. Jena, Germany: Gustav Fischer.
Bortolotti, L. (2009). The epistemic benefits of reason giving. *Theory and Psychology, 19*(5), 624–645.
Bortolotti, L. (2015). The epistemic innocence of motivated delusions. *Consciousness and Cognition, 33*, 490–499.
Bortolotti, L., & Cox, R. E. (2009). "Faultless" ignorance: Strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition, 18*, 952–965.
Carson, T. L. (2010). *Lying and deception: Theory and practice*. Oxford: Oxford University Press.
Cooper, J., & Carlsmith, K. M. (2001). Cognitive dissonance. *International Encyclopedia of the Social and Behavioural Sciences*, 2212–2214.
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*(6), 1314–1329.
Dalla Barba, G. (1993). Different patterns of confabulation. *Cortex, 29*, 567–581.
De Houwer, J., Teige-Mocigemba, K., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*(3), 347–368.
Duncan, B. L. (1976). Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping blacks. *Journal of Personality and Social Psychology, 34*(4), 590–598.
Fallis, D. (2009). What is lying? *The Journal of Philosophy, 106*, 29–56.
Fine, C. (2010). *Delusions of gender*. UK: Icon Books Ltd.
Fotopoulou, A. (2008). False selves in neuropsychological rehabilitation: The challenge of confabulation. *Neuropsychological Rehabilitation, 18*(5–6), 541–565.
Fotopoulou, A., Conway, M. A., & Solms, M. (2007). Confabulation: Motivated reality monitoring. *Neuropsychologia, 45*, 2180–2190.
Fotopoulou, A., Conway, M. A., Tyler, S., Birchall, D., Griffiths, P., & Solms, M. (2008). Is the content of confabulation positive? An experimental study. *Cortex, 44*, 764–772.
Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "Implicit" attitudes unconscious? *Consciousness and Cognition, 15*, 485–499.
Gendler, T. (2011). On the epistemic costs of implicit bias. *Philosophical Studies, 156*, 33–63.
Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*(3), 1369–1392.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.
Hall, L., & Johansson, P. (2008). Using choice blindness to study decision making and introspection. In Gardenfors, P., & Wallin, A. (Eds.), *Cognition – A smorgasbord* (pp. 267–83).
Heilman, M. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues, 57*(4), 657–674.
Hirstein, W. (2000). Self-deception and confabulation. *Philosophy of Science, 67*(Suppl.), S418–S429 (Proceedings of the 1998 biennial meetings of the philosophy of science association. Part II: Symposia papers).
Hirstein, W. (2005). *Brain fiction: Self-deception and the riddle of confabulation*. MIT Press.
Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Psychology, 43*(3), 274–306.

Holroyd, J. (2015). Implicit bias, awareness and imperfect cognitions. *Consciousness and Cognition, 33*, 511–523.
Kennett, J., & Fine, C. (2009). Will the real moral judgment please stand up? *Ethical Theory and Moral Practice, 12*(1), 77–96.
Kopelman, M. D. (1999). Varieties of false memory. *Cognitive Neuropsychology, 16*(3/4/5), 197–214.
Lackey, J. (2013). Lies and deception: An unhappy divorce. *Analysis, 73*(2), 236–248.
McKay, R., & Kinsbourne, M. (2010). Confabulation, delusion, and anosognosia: Motivational factors and false claims. *Cognitive Neuropsychiatry, 15*(1/2/3), 288–318.
Metcalf, K., Langdon, R., & Coltheart, M. (2010). The role of personal biases in the explanation of confabulation. *Cognitive Neuropsychiatry, 15*(1–3), 64–94.
Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition, 19*(4), 395–417.
Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *PNAS, 109*(41), 16474–16479.
Nier, J. A. (2001). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Process & Intergroup Relations, 8*(1), 39–52.
Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., et al (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*(1), 36–88.
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*(2), 181–192.
Ramachandran, V. S. (1996). The evolutionary biology of self-deception, laughter, dreaming and depression: Some clues from anosognosia. *Medial Hypotheses, 47*, 347–362.
Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology, 39*, 590–598.
Saul, J. (2012a). Ranking exercises in philosophy and implicit bias. *Journal of Social Philosophy, 43*(3), 256–273.
Saul, J. (2012b). Scepticism and implicit bias. *Disputatio, 5*(37), 243–263.
Sorenson, R. (2007). Bald-faced lies! Lying without the intent to deceive. *Pacific Philosophical Quarterly, 88*, 251–264.
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*(7/8), 509–528.
Van Leeuwen, Neil D. S. (2007). The spandrels of self-deception: Prospects for a biological theory of a mental phenomenon. *Philosophical Psychology, 20*(3), 329–348.
Van Damme, I., & d'Ydewalle, G. (2010). Confabulation versus experimentally induced false memories in Korsakoff patients. *Journal of Neuropsychology, 2*(2), 211–230.
Zangwill, O. L. (1953). Disorientation for age. *The British Journal of Psychiatry, 99*(417), 698–701.