Josefa Toribio

ICREA-UB [1, 2]

[1] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain.
[2] Universitat de Barcelona. Department of Philosophy, Montalegre 6, Barcelona 08001, Spain

email: jtoribio@icrea.cat

Phone: +34 934 037991
Fax: +34 934 037980

Acknowledgements: A version of this paper was presented at the Institute of Philosophy (London) as part of their Logic, Epistemology and Metaphysics Seminar Series. I would like to thank the audience there, especially Conor McHugh, for their helpful comments. Many thanks to Indrek Reiland, who was kind enough to help me think through one of the main objections raised by the referees of this journal. My thanks also go to them.

Funding: Research for this paper was supported by the MINECO (Ministerio de Economía y Competitividad) via research grant MCINN FFI2014-51811, by the EC, Project: 675415 – DIAPHORA, H2020-MSCA-ITN-2015, and by AGAUR (Agència de Gestió d'Ajuts Universitaris i de Recerca) via research grant 2017-SGR-63.

Accessibility, implicit bias, and epistemic justification

Josefa Toribio
ICREA-UB

Abstract

It has recently been argued that beliefs formed on the basis of implicit biases pose a challenge for accessibilism, since implicit biases are consciously inaccessible, yet they seem to be relevant to epistemic justification. Recent empirical evidence suggests, however, that while we may typically lack conscious access to the source of implicit attitudes and their impact on our beliefs and behaviour, we do have access to their content. In this paper, I discuss the notion of accessibility required for this argument to work vis-à-vis these empirical results and offer two ways in which the accessibilist could meet the challenge posed by implicit biases. Ultimately both strategies fail, but the way in which they do, I conclude, reveals something general and important about our epistemic obligations and about the intuitions that inform the role of implicit biases in accessibilist justification.

Keywords: accessibilism; implicit bias; propositional justification; conscious access

1. Introduction

Accessibilism is the view that only consciously accessible factors are relevant to epistemic justification. It has been recently argued (Puddifoot, 2016) that the justification of beliefs formed indirectly as a result of implicit biases, i.e., beliefs formed as the result of how things seem to us given implicit biases' influence on the available evidence, pose a problem for accessibilism—since implicit biases are consciously inaccessible, yet they seem to be relevant to epistemic justification. In this paper, I set out to do three things. First, I focus on how best to understand the way in which implicit biases are said to be inaccessible. I do this by reviewing some recent empirical evidence which suggests that, while we may typically lack conscious access to the source of implicit attitudes and their impact on our beliefs and behaviour, we do have access to their content (Gawronski et al. 2006; Hall & Payne 2010). Second, I discuss the notion of accessibility required for Puddifoot's argument to work in light of the reviewed empirical evidence and argue that accessibilism could meet the challenge posed by implicit biases in at least two ways. Finally, I show that these versions of accessibilism only get us out of the implicit bias challenge by positing an implausibly over-intellectualized and over-reflective subject. Although ultimately both strategies fail, the way in which they do, I conclude, reveals something general

and important about our epistemic obligations and the role of implicit biases in (accessibilist) justification.

2. Accessibilism

Accessibilism is a variety of internalism. According to accessibilism, whether or not a belief is justified depends solely on factors that are consciously accessible to the subject—typically, the mental states a subject can reflect about (see e.g. BonJour 1980; Chisholm 1988 or Steup 1999). Accessibilism is thus different from mentalism, another variety of internalism, according to which the only factors that determine justification are the subject's mental states, regardless of whether they are consciously accessible or not (see e.g. Conee and Feldman 2004; Feldman, 2005).

Accessibilism takes two different forms depending on which internal factors are taken to be relevant to justification. A weak form of accessibilism requires having access just to the belief's justifiers. For my belief that e.g. Clara is wearing a black shirt to be justified, a perceptual experience of Clara wearing a black shirt would be such a justifier. Modulo defeaters—for which the accessibility constraint also applies—a perceptual experience of this kind is considered my reason for believing that Clara is wearing a black shirt. The fact that I undergo such a perceptual experience is a reason I can take into account in any thoughts involving my belief. Accessibilism of this kind is the thesis that the justification of a belief *p* supervenes upon facts that the subject is able to know by reflection alone. On some versions of accessibilism, what we can know by reflection alone also includes *a priori* knowledge and memory of all knowledge thus acquired.[1]

A strong form of accessibilism holds that, for a proposition *p* to be justified, we must also be able to consciously access *p*'s justificatory status, i.e., we need to be aware that *p*'s justifiers justify *p* (see e.g. BonJour, 1985, ch. 2). When considering the belief that Clara is wearing a black shirt, the strong accessibilist thus requires that I am aware that my perceptual experience as of Clara wearing a black shirt justifies my belief that Clara is wearing a black shirt, i.e., I have to be aware of my experience

[1] Pryor (2001, p. 104) labels this view 'simple internalism'.

as being the reason for my believing what I do. What both forms of accessibilism have in common is a commitment to the view that only consciously accessible factors are directly relevant to epistemic justification.[2]

Accessibilism is a thesis about propositional justification, i.e., about which propositions a subject is justified to believe given the available evidence. It is not a thesis about doxastic justification, i.e., about a subject's justifiably believing what she does. One can have good reasons *R*, and thus be propositionally justified to believe *p*, even if one does not believe it or believe it for reasons other than *R*. To justifiably believe *p*, i.e., to be (doxastically) justified to believe what one does believe, requires propositional justification plus some additional grounding or causal connection between what one believes and the reasons for believing it. The distinction between propositional and doxastic justification is important so as to avoid unwarranted objections to accessibilism.

The deontological view of epistemic justification is often cited as one of the central motivations for accessibilism. According to this view of justification, one is justified to believe just in case one has flouted no epistemic obligations in the pursuit of true beliefs. In turn, it is often argued, we can fulfil our epistemic duties to obtain true beliefs just in case we are aware of the reasons we have to believe what we do. We are blameless to believe *p* just in case what justifies *p* is accessible to us, i.e. just in case we have a reason to believe *p*. Only what is accessible can, on this view, act as reason. Whether a subject is justified to believe *p* thus supervenes on what is accessible to her.

Other motivations for accessibilism rest on similar intuitions about the need for the subject to be consciously aware of what counts as justifiers of her beliefs, if they are to be justified. BonJour's (1980) classic case of Norman, the clairvoyant, whose belief that the President is in NYC is nomologically linked to the presence of the President in NYC, is supposed to make this idea vivid: Norman's reliability falls short of justifying his belief because the fact that he is reliable is not accessible to him.

---

[2] I ignore here different versions of what is considered to be the appropriate kind of justifiers within each type of accessibilism. For instance, not every accessibilist would agree that perceptual experiences themselves, as opposed to the beliefs based on perceptual experiences, count as justifiers for other beliefs. These details are not important for the discussion that follows.

What is known as *the new evil demon problem* for reliabilism (Lehrer and Cohen, 1983; Cohen, 1984) has also been used to elicit accessibilist intuitions about justification. In new evil demon scenarios, we assume that most of our beliefs in the actual world are the result of reliable mechanisms. We then imagine a possible world in which we form exactly the same beliefs as in the actual world, based on exactly the same kind of experiences and through exactly the same reasoning processes. However, in that possible world, an evil demon makes sure that our experiences systematically deceive us by making our otherwise reliable cognitive mechanisms unreliable—the evil demon creates non-veridical perceptual experiences that are qualitatively identical to our veridical ones in the actual world. If reliabilism is true, our beliefs will be justified while our evil demon world counterparts' beliefs will not, since our beliefs are formed reliably and the evil demon world people's beliefs are formed unreliably. Yet this result is supposed to strike us as counterintuitive, for the available evidence is exactly the same in both the real and the evil demon world.

Traditionally, what can be accessed and hence the reasons people have to believe are taken to be only facts that we are in a position to know by reflection alone. However, it has been recently argued (see e.g. Gibbons, 2006 and Hatcher, 2016) that what is required for appropriately holding a subject epistemically responsible may also be reasons that are "easily knowable" (Hatcher, 2016, p. 17). The idea, to borrow Gibbons' (2006, p. 36) phrase, is that "justification supervenes on what you are in a position to know", where what you are in a position to know are facts you ought to know, given the epistemic situation in which you are, even if you are not aware of such facts by reflection alone. I return to this in Section 6.[3]

3. Implicit bias and accessibilism

Implicit biases or implicit attitudes (henceforth, I take the two expressions to be

[3] It has been argued (Hatcher, 2016, ft. 39) that Gibbons' proposal cannot capture the intuition prompted by new the evil demon scenario because facts about which things are easily knowable are different in our world and in the evil demon world. It would thus be false that both our beliefs and our evil demon world counterparts' are equally justified. Be this as it may, the issue does not affect my argument about the challenge of implicit biases to accessibilism. In fact, it does help fine-tune one of my proposals. See below.

synonymous) are representational mental states that reflect stereotypical properties of members of, and items in, all kinds of different categories: racial groups, professions, women, nationalities, members of the LGBTQ community, moral and political values, etc. They typically connect one or two concepts and a valence (either negative or positive) or two or more concepts, one of which has either a negative or a positive slant. There is no general agreement about the representational structure of implicit biases. Some philosophers take them to be *sui generis* mental states with an associative structure (see e.g. Gendler 2008a,b; 2011; Brownstein & Madva 2012; Madva 2016), while others view them as just plain beliefs (De Houwer 2014; Egan 2011; Hughes et al. 2011; Mandelbaum 2016; Mitchell et al. 2009; Smith 2012), or as states that fall short of being beliefs but are nevertheless propositionally structured (Levy 2014). Although only on some of these views implicit attitudes are characterized as unconscious (Mandelbaum 2016) and although this claim has recently come under attack in social psychology (see e.g. Hahn et al. 2014 and Section 4 below), it is still quite common in social psychology to find lack of introspective awareness as a distinctive feature of these representational states.[4] Standard social psychology textbooks, for instance, describe implicit attitudes as unconscious attitudes that we "cannot self-report in questionnaires because we are not aware of having them" (Kassin et al. 2010, p. 207; see also Kenrick et al. 2010). Often in the literature, 'unconscious' and 'implicit' are used interchangeably (see e.g. Cunningham et al. 2004; Quillian 2008), with some social psychologists explicitly holding the view that implicit attitudes just cannot be introspected (see e.g. Greenwald & Banaji 1995; McConnell et al. 2011).

Especially when considering implicit biases such as racism, sexism or homophobia, the central idea seems to be that, despise sincerely and justifiably considering ourselves to be unprejudiced agents, consciously committed to egalitarianism in all its forms, we are often surprised to discover that we still harbour implicit attitudes that betray our unprejudiced, egalitarian explicit views. This kind of mismatch between our explicit and our implicit attitudes is often used to argue that implicit biases are unconscious. The assumption that implicit attitudes are

[4] There are also philosophers who argue that all our attitudes, both explicit and implicit, are unconscious (see e.g. Carruthers 2017; King & Carruthers 2012). If this is true, the challenge to accessibilism will not be confined to implicit biases. The discussion of this topic is beyond the scope of this paper, but if my argument here works for implicit biases, then it will also generalize to accommodate this view.

unconscious mental states whose content can be diametrically different to the content of our explicit (self-reported) ones also makes it plausible to think that only indirect methods would give us information about them. The Implicit Association Test (IAT) (Greenwald et al., 1998) and sequential priming, together with other tests,[5], have thus become classic tools for unmasking the degree to which we are subject to the tyranny of such biases and are widely used in Social Psychology. For instance, in Keith Payne's (2001) now classic weapon identification task, it is shown that participants identify weapons much faster when primed with pictures of black faces compared to pictures of white faces. Participants are also more prone to misidentify tools as weapons when primed with pictures of black faces as opposed to pictures of white faces.

In order to highlight how accessibilism seems to deliver the wrong verdict when accounting for the justification of beliefs formed indirectly as a result of implicit biases, Katherine Puddifoot (2016) asks us to imagine two different scenarios. In the first one, Jones, a member of a jury in a rape case involving a black man, considers all available evidence provided by the prosecution and finds it convincing that the defendant is guilty. Not just him, all other members of Jones' community also find that the evidence strongly supports the belief that the black man is guilty. Jones thus has good reasons to believe that the defendant is guilty and believes that he is guilty for those reasons. In the second scenario, the evidence remains the same and so do the opinions of other members of the community, but here, for both Jones and the members of his community, the evidence seems convincing only because they hold an implicit bias against black men. In the second scenario, Jones associates black men with violence (or believes that black men are violent, if the propositional model is your preferred model) and is "generally more incredulous" (p. 422), says Puddifoot, so, were not for his implicit racist attitude, the available evidence would not seem compelling to him. In this second scenario, Puddifoot claims (2016, p. 422), Jones' belief that the defendant is guilty is not justified or, at a minimum, its justificatory status should strike us as much weaker. Puddifoot relies on this pre-theoretical intuition to argue against (both forms of) accessibilism. Here is her argument

[5] E.g., the Affect Misattribution Procedure (AMP) (Payne et al. 2005) or the Go/No-go Association Task (GNAT) (Nosek & Banaji 2001).

(Puddifoot, 2016, p. 422. ACCESS henceforth):

ACCESS:

1. According to accessibilism, only consciously accessible factors can be relevant to epistemic justification.
2. Implicit biases are consciously *in*accessible factors.
3. Implicit biases are relevant to epistemic justification.
   Therefore:
4. There are some consciously *in*accessible factors that are relevant to epistemic justification.
5. Accessibilism is wrong.

Not everybody will feel the pull of the intuition behind Puddifoot's argument. In particular, one may wonder whether the details of the scenario really target accessibilism. For accessibilism is a view about propositional—not doxastic—justification. Yet, Puddifoot's case stipulates that the available evidence is exactly the same for Jones both in scenario 1 and scenario 2. It will thus follow that Jones' beliefs in both scenarios have the same level of propositional justification. The only difference between scenario 1 and 2 are the reasons for which Jones and his counterpart believe what they do, so the difference seems to be a difference in doxastic justification. Jones 1 is, while Jones 2 is not doxastically justified in believing what they do.[6]

Even with this important proviso, there is something about the intuition behind the spirit, if not the letter, of Puddifoot's argument that I think is worth discussing as an argument against accessibilism. After all, accessibilism still requires that the reasons for believing *p* must all be reasons that we are aware of, and does so motivated, in part, by a deontological view of epistemic justification. The type of scenario suggested by Puddifoot's case strikes us as initially plausible as a case against accessibilism (properly understood), if it does, because we feel that Jones in scenario

[6] Conor McHugh (in conversation) raises a further concern. It is not even completely clear, he claims, that Jones 1 and 2 differ with regard to whether his beliefs are doxastically justified. Although Jones' implicit racism in scenario 2 affects his assessment of the evidence, this influence by a variable irrelevant to truth still makes him reach the right conclusion since, by stipulation, that the defendant is guilty is justified by the available evidence. So, the details of the thought experiment would have to be much more elaborated to even get a difference in doxastic justification. I intentionally and charitably overlook this problem as well as the more important issue of Puddifoot's argument failing to address propositional justification. See below.

2 fails to fulfil some epistemic duty—even if, to do so, it he would have to be aware of something he is not. So I assume that we could tweak Puddifoot's case enough to make it problematic for accessibilism.[7] This is how I will proceed from here. I show that recent research in social psychology warns us against the widespread conception of implicit biases as unconscious when thinking about their content. I rely on this research to reject premise 2 in ACCESS, but offer instead a refined version of the argument, ACCESS_2 (Section 5). I then argue (Section 6) that the accessibilist can still meet the challenge posed by ACCESS_2—or so it seems. Ultimately, I seek to debunk the intuition behind ACCESS_2 (Section 7), but, hopefully, we would have learnt a lot about accessibilism, epistemic responsibility and implicit biases on our way to the final conclusion.

4. Implicit bias: the evidence. Unaware of what?

Researchers in social psychology have become increasingly interested in whether the frequently observed gap between explicit (self-reported) and implicit (indirect) attitude measures should be taken to straightforwardly reflect a distinction between conscious and unconscious attitudes. When looking for empirical evidence, Gawronski et al. (2006) claim, we should keep in mind that there are three different ways in which we can say of an attitude that it is unconscious. 'Unconscious' may refer to the lack of awareness we have of the origin of our attitudes; what they call *source aw*areness. 'Unconscious' may refer instead to our lack of awareness of the content of the attitude: *content awareness*. Finally, when charactering implicitly held attitudes as unconscious, we may want to refer to our failing to be aware of their impact on other mental states, psychological processes or behaviour: *impact awareness* (Gawronski et al., 2006, p. 486). These three dimensions of unconsciousness are logically related. Without being aware of the content of an attitude, we could not be aware of its source or its impact. So, content awareness is

[7] Perhaps, we could add that, in scenario 1, Jones is excessively credulous or just more credulous than in scenario 2. As I pointed out earlier, Puddifoot does mention in passing that being generally more incredulous in scenario 2 may be why, were not for the influence of his implicit racist bias, Jones would not find the available evidence convincing. Or perhaps Jones, in scenario 2, is less attentive than he is, in scenario 1, to what is exactly the same available evidence from the point of view of the evidence relevant for propositional justification so that, again, were not for the bias, Jones would fail to be convinced by it.

necessary for source and impact awareness, but it is not sufficient. We may be aware of the content of an attitude without being aware of how we acquired it or how it affects other of our mental states and psychological processes.

Gawronski at al. (2006) meta-analysis of a variety of studies about the three related dimensions of unawareness leads to a triple conclusion. First, it is fairly common to lack awareness of the origin of our attitudes. However, such lack of source awareness is not a distinctive mark of implicit bias, since it also affects our explicit attitudes. Second, and surprisingly in contrast to the prevalent view, the studies show that we are often aware of the content of our implicit bias. Lack of awareness of the content of our attitudes is typically inferred from low correlation between self-reported attitudes and those that emerge from indirect measures. Yet, there is now growing evidence that the gap between implicit and explicit attitudes is often due to factors other than our being unaware of the former's content—cognitive, motivational and methodological factors. Hall and Payne (2010) thorough meta-analysis of racial biases also favours the hypothesis that what best explains the low correlation between implicit and self-reported attitude measures is not lack of content awareness, but people's reluctance to openly report their own racial biases. This is more clearly so in the case of highly reflective subjects. Both meta-analyses, Gawronski at al. (2006) and Hall and Payne (2010) refer to a study by Nier (2005) in which he used the so-called "bogus pipeline" manipulation, i.e., letting some of the participants believe that the experimenters could always detect whether their racial attitudes as measured by self-reported evaluations were accurate. The correlation between implicit and explicit attitudes was much higher in the group of participants make to believe this, thus showing that cognitive and motivational factors about presenting themselves as less racist individuals in self-reported evaluations are behind typical lower correlation results—not lack of awareness of the content of their implicit attitudes.

The correlation between implicit and explicit attitudes can also fluctuate depending on whether the measure of explicit attitudes involves affective as compared to cognitive elements. Gawronski and collaborators (2006) report a couple of studies (Banse et al., 2001; Hofmann et al. 2005) in which the correlation between self-reported and implicit measures of attitudes toward homosexuals is much higher when the self-reports involve affective reactions than when they involve general

descriptions. In other words, the content of people's implicit homophobic attitudes seems to be much more content-conscious when testing for it involves descriptions about feelings (e.g. how subjects feel about witnessing certain sexual encounters between people of the same sex) than when testing involves opinions or general views about homosexuality.

Hall and Payne (2010) also isolate a similar factor that explains low correlations between implicit and explicit attitude measures better than lack of content awareness, namely, the fact that subjects tend to be confused about what they should consider an attitude in the first place, so they self-reported views are often skewed. They review a study by Ranganath et al. (2008) that makes it clear, for instance, that when subjects are experimentally forced to take their gut reactions toward gay people as indicators of their attitudes, the gap between implicit and explicit attitude measures is narrower. The study suggests "that subjects have some awareness of the attitudes revealed by implicit tests because when asked the 'right' questions, they can report them in a way that matches their responses on implicit tests" (Hall and Payne, 2010, p. 227).[8]

Finally, both Gawronski et al. (2006) and Hall and Payne's (2010) meta-analyses provide empirical evidence about certain methodological flaws on the measurement of implicit attitudes, which seem, again, to better explain low correlations between implicit and explicit attitudes than lack of content awareness. A common flaw is the lack of internal consistency among different implicit attitude measures. Most methods for evaluating implicit attitudes rely on response latencies, which exhibit a high rate of measurement error.[9] When studies are designed in such a way so as to control for

[8] These results are not a knockdown argument against the inaccessibility of the content of our implicit biases. They only suggest that we are more aware of their content than previously assumed. It's just that social psychologists have been asking the wrong sort of questions. Whether or not subjects need to be cognitively sophisticated to have introspective access to the content of their implicit attitudes is a thorny issue. Hahn et al.'s previously mentioned (2014) study shows, on the one hand, that fairly cognitively unsophisticated subjects are really good at predicting their own performance on the IAT across different experimental conditions, even when they are told very little about the test or about what implicit attitudes are, thus reinforcing the view that our awareness of the content of implicit attitudes is greater than formerly thought regardless of participants' cognitive sophistication. On the other hand, it could be argued that there may be some implicit-attitude-relevant but subtle questions that only cognitively sophisticated subjects can really ask *themselves*, i.e., outside experimental settings. The issue of cognitive sophistication will play an important role in the final part of my argument. See Section 7.

[9] See, in particular, the recent controversy over the studies that link subjects' IAT scores and their actual discriminatory behaviour. Greenwald, Poehlman, Ulmann and Banaji (2009) argue for a strong link between these two variables. Oswald, Mitchell, Blanton, Jaccard and Tetlock (2013) question the

measurement error, the gap between explicit and implicit attitudes measures is, again, narrower. The same occurs when experimenters use methods that do not rely on response latencies, such as the Affect Misattribution Procedure (AMP), which also exhibits a high reliability and high internal consistency (Hall and Payne, 2010, p. 226).

The typically assumed hypothesis that implicit attitudes are unconscious mental states, in the sense of our not being aware of their content, thus loses plausibility when all these different variables are taken into account. Lack of *impact* awareness, by contrast, becomes the key issue in this discussion. Both Gawronski et al. (2006) and Hall and Payne (2010) meta-analyses highlight this point. When looking at the evidence, what seems to be widely confirmed is that subjects are not aware of the influence that their implicit attitudes have on their other mental states, psychological processes and behaviour, even when they are aware of their content, are motivated to control for their influence and have enough cognitive capacity to do so (Gawronski et al., 2006, p. 491).[10]

In a couple of studies involving a simple memory task, Payne and collaborators (Payne et al. 2004) examine the contrast between participants' subjective experience about the influence of a racist bias in their pairing of stereotypical black and white names with stereotypical black and white occupations (basketball player and politician, respectively) with the actual demonstration of the bias. First they ask participants to memorize a list of names paired with one of these two occupations, some of which are consistent with the stereotype and some of which are not. Then

---

link and focus on the influence of overt biases in the participants. Greenwald, Banaji and Nosek (2015) quickly replied to the Oswald et al. meta-analysis. Additional studies since then keep feeding the debate.

[10] As an anonymous referee points out, lack of impact awareness, like lack of source awareness, is a property that affects both implicit and explicit attitudes. Explicit attitudes, such as explicit beliefs, desires or fears often have all sorts of unknown effects on other mental states and behaviour. So, if the challenge to accessibilism stems from our generally being unaware of the impact of implicit biases on thought and behaviour, the same will apply when considering explicit attitudes. It is revealing that one of the main conclusions of Hall & Payne's (2010) meta-analysis is that "an attitude need not be unconscious to influence our thoughts and behaviors without our awareness" (p. 229). Again, discussion of this topic goes beyond the scope of this paper. I contend, however, that my argument about implicit biases vis-à-vis accessibilism will successfully generalize to cover the unbeknownst effects on thought and behaviour of the relevant explicit attitudes.

they ask participants to recall the occupation each of the names was paired with and also, and importantly, how confident they are that their answer is correct. The study shows that when participants could remember the pairs correctly, correlation between confidence in correctness and real correctness was high. But this process was, of course, controlled by memory. When memory failed and the recalling process reflected automatic processing, participants were often wrong, i.e., they misremembered which name was paired with which occupation. They were also more likely to pair stereotypical black names with the occupation of basketball player and stereotypical white names with politician. Interestingly, and relevant for my purposes here, they manifested this bias while both reporting perfect confidence and no confidence at all in their memories (Hall and Payne, 2010, p. 231).

If it turns out, as these results suggest, that we are, for the most part, aware of the content of our implicit biases, even if their impact on other mental states and behaviour is not consciously accessible, does ACCESS lose much of its force? I turn to this issue in the next Section.

5. The challenge of implicit bias to accessibilism. First Pass

On standard characterizations of accessibilism, the relevant accessible justifiers are always contentful states or epistemic standards: beliefs, experiences and the like. Here is e.g. Matthias Steup's (1996, p. 84) classic formulation:

> What makes an account of justification internalist is that it imposes a certain condition on those factors that determine whether a belief is justified. Such factors—let's call them "J-factors"—can be beliefs, experiences, or epistemic standards. The condition in question requires J-factors to be *internal to the subject's mind* or, to put it differently, *accessible on reflection*.

'Consciously accessible factors' in premise 1 of ACCESS refers to contentful mental states, i.e., the truth of accessibilism entails that only consciously accessible contentful mental states can be relevant to epistemic justification. Yet, the results we reviewed in the previous Section reveal that—especially with enough motivation and cognitive capacity—we are often aware of the content of our implicit attitudes. If so, premise 2 is false and accessibilism gets easily off the hook.

This would be too quick though, for at least the two following reasons. First,

research in social psychology does not completely rule out lack of content awareness with regard to implicit biases. It simply acknowledges that the mismatch between our biases and our explicit attitudes is much narrower than it is often assumed, thus suggesting that it is a mistake to talk loosely about (the content of) implicit bias as unconscious. As Puddifoot (2016, ft. 3) points out, all that is needed for ACCESS to work is that implicit biases, whose content is consciously inaccessible, are, at least sometimes, relevant to epistemic justification.[11] It could thus be argued that the situation in scenario 2 is widespread enough so as to make trouble for accessibilism (but see below).

Second, Puddifoot's discussion throughout the paper is often phrased as if the consciously inaccessible—yet justificatorily relevant factors—are not the attitudes themselves, but the *influence* of implicit attitudes on thought. This certainly is in agreement with the results from social psychology: we, as a rule, lack awareness of the impact that implicit biases have on the rest of our mental life. The force of the intuition about the different justificatory status of Jones' belief in scenario 2 thus seems to come from the fact that, in such a set-up, the available evidence seems convincing to Jones only as a result of the *influence* of his racial bias on his decision making—an influence he is not aware of. ACCESS is thus best formulated as ACCESS_2:

1'. According to accessibilism, only consciously accessible factors can be relevant to epistemic justification.
2'. The impact of implicit biases on other cognitive states is a consciously *in*accessible factor.
3'. The impact of implicit biases on other cognitive states is relevant to epistemic justification.

Therefore:

4'. There are some consciously *in*accessible factors that are relevant to epistemic justification.
5'. Accessibilism is wrong.

The accessibilist could reply, however, that there is an equivocation on 'factor' in

[11] Although, arguably, her argument would be much weaker if it turned out that we are aware of the content of our implicit biases most of the time.

ACCESS_2. In premise 1', 'factor' refers to contentful mental states, while 'factor' refers to a causal influence between states in the rest of the argument. In other words, the accessibilist could just deny premise 3'. After all, premise 3' seems to beg the question against accessibilism, since the influence of implicit biases on other mental states—understood as a causal influence—would be an obviously inaccessible factor. Puddifoot acknowledges (a version of) this possible rejoinder, but argues that denying premise 3' is inconsistent with central motivations for accessibilism. Her argument thus has the form of a dilemma. Either the accessibilist accepts that the impact of implicit biases on thought is relevant for epistemic justification, thus denying accessibilism's core view (premise 1'), or she denies such relevance (premise 3'), undermining as a result some of the most important motivations for holding an accessibilist position in the first place (Puddifoot, 2016, p. 423).

One of these motivations is the commitment to a deontological view of epistemic justification, i.e., the commitment to the idea that being justified in believing something is essentially linked to a believer's duty to take all necessary steps to avoid falsehood. Yet, according to Puddifoot, being thus motivated by such a commitment delivers, again, the wrong verdict with regard to scenario 2. Jones, in this scenario, has done all there is in his power to obtain a true belief: he has fulfilled all his epistemic responsibilities and considered all available evidence as well as the opinions of other members in his community. Since *Ought implies Can*, and causal influences between mental states are not among the J-factors over which Jones can have any responsibility, he should be in the clear when holding the belief that the defendant is guilty. This is, however, counterintuitive because the example, Puddifoot notes, forces us to acknowledge that there is a difference between scenario 1 and 2 with regard to the justificatory status of Jones' belief.

To sum up. Even if we have conscious access to the content of our implicit biases, this does not entail that we also have conscious access to their impact on our thought and behaviour. Yet, when reformulating ACCESS in terms of impact accessibility instead of content accessibility, accessibilism still seems to deliver the wrong verdict with regard to the justification of beliefs formed indirectly as the result of our implicit biases' influence. So, ACCESS_2 retains a certain appeal.

In the next Section, I will put forward two different responses the accessibilist can

offer to meet the implicit bias challenge. Both of them support the connection between accessibilism and a deontological view of justification.

6. The challenge of implicit bias to accessibilism. Second Pass

As we saw, accessibilism is a supervenience thesis. The justification of a belief supervenes upon what is consciously accessible to the believer. Accordingly, something can be a justifier, and hence a reason, just in case it is knowable by the believer. This formulation dovetails with the deontological view of justification, since, as I pointed out above, subjects can meet their epistemic obligations only if they can come to know them. However, the notion of accessibility is consistent with a wide reading of what is know*able*. Epistemic responsibility and accessibility need not be and should not be restricted just to what we know. Depending on our epistemic situation, we may be responsible for things that we do not know, but that we are in a position to know. It could thus be argued that even if the impact of implicit biases—as a causal influence on thought—is not something subjects can be introspectively aware of, the evidence for and the beliefs about such a causal impact are knowable factors in this sense: they are factors subjects are in a position to know and hence ought to know. The first move I want to make for getting accessibilism to meet the implicit bias challenge is thus to widen the supervenience base for justification in such a way so as to include knowledge about the pervasiveness of the impact of implicit biases on our thinking, decisions and behaviour.

The notion of being in a position to know has by now great philosophical pedigree, even if it started as a central part in anti-luminosity arguments (Williamson, 2000, ch. 4). My usage here relies heavily on John Gibbons' (2006) treatment of the formula in his argument in favour of what he calls 'access externalism'. Gibbons' label could be misleading, since my target is to make *accessibilism* meet the challenge of implicit biases. I hope to dispel any concerns about this matter in what follows.

Gibbons' key move is to divorce the notions of accessible and internal in such a way so as to allow for (some, but not all) external facts to be accessible. This is achieved by understanding the notion of accessible fact as facts that one is in a position to know—instead of facts that one can know by reflection alone. On

Gibbons' proposal, the facts that one is in a position to know may thus include some external facts, which will be different in different epistemic situations. Gibbons motivates the sort of considerations that lead him to his revised notion of accessibility with an example. Someone—let's call him John—forms the belief that he is going to have a jalapeño, mushroom and cream cheese omelette for breakfast after carefully checking the night before that all necessary ingredients are in the fridge and knowing that, as a matter of fact, his partner hardly ever eats breakfast. While getting the ingredients ready in the morning, John believes that he will soon have a jalapeño, mushroom and cream cheese omelette. However, unnoticed by him, there is a note stuck to the fridge that says: "We are out of cream cheese". Importantly, to get the story right, we are supposed to imagine that it is customary in John's household to leave notes of this kind on the fridge, so that even though John had not noticed the note, he should have. And if John should have noticed the note but has not, then his belief about what he is going to have for breakfast is, on Gibbons' account, not justified.

The important point is that the note on the fridge—and not just what is introspectively accessible to John—makes all the difference for the justification of John's belief. It makes all the difference because, although it is an external factor, John is in a position to know about the existence of this type of note. It is John's obligation to check for them when forming beliefs about what he is going to have for breakfast. Of course, not all unnoticed evidence destroys justification and not all external facts are thus accessible and hence relevant for justification. Only unnoticed evidence that subjects are in a position to know, given the epistemic situation they are in. Here is Gibbons' contrasting scenario to clarify the distinction. Imagine that, instead of sticking the note on the fridge, the household member who wrote it put it, absentmindedly, in one of her pockets. In this case, the evidence is there, but John is not in a position to know about it. This fact is not accessible to him even on this understanding of accessibility. According to Gibbons, in this second scenario, John's belief about what he is going to have for breakfast is justified.

Jones, in Puddifoot's second scenario, does not have introspective access to the impact of his racist bias on the assessment of the evidence that results in his belief about the defendant. Yet, racism, and its influence on our thinking is a pervasive fact.

It is not, as it were, a piece of information hidden from members of our community; it is as easily knowable as Gibbons' note on the fridge. It is the kind of fact that helps configure Jones' epistemic situation. Jones ought to know about such a fact, and if Jones ought to know about such a fact, then he can know about it, which is precisely what we want to say when we say that he is in a position to know.

What establishes the boundaries of the extended supervenience base for justification proposed by Gibbons with regard to which external facts are relevant for justification is linked, in this way, to what we can reasonably hold Jones responsible for, given the epistemic situation he is in. In a situation like a trial, where the standards for justification are particularly high, and given the pervasiveness of implicit biases and their influence on our thinking, Jones' failing to take into account their (potential) influence when carefully scrutinizing the available evidence about the black defendant puts him in the wrong from the "wide" accessibilist point of view recommended here.[12] His belief is not justified. The impact of implicit biases on our thinking, decisions and behaviour should be treated as the customary note on the fridge that we all ought to check whenever forming beliefs where such an impact is highly likely. We ought to know about such facts, and we ought to know only if we can know, i.e., only if we are in a position to know—as the formula is understood here. This version of accessibilism thus meets the implicit bias challenge while suitably responding to the demands of the deontological notion of justification.

For those who may still harbour some suspicion that *wide* accessibilism is an undercover form of externalism, as the title of Gibbons' paper invites to think, let me try a different move. This second strategy does not require widening the supervenience base of justification. Instead, we need to pay attention to exactly those facts to which Jones has introspective access. To do that, I would like to distinguish between the *very impact* of Jones' racist bias on his belief about the black defendant and the fact that such impact is (or is not) accessible to him. We can grant that the very impact of Jones' racist bias on his belief is not accessible to him while accepting that he has access to the fact that such an impact is not accessible. This is, after all, why social psychologists can set up experimental conditions to check for implicit biases impact awareness. Jones indeed does not have introspective access to the

[12] As labels go, I prefer the label *wide accessibilism* to *access externalism*. The proposal remains faithful to Gibbons' formula though.

influence of his racist views on his final verdict. Yet, like most of us, Jones has introspective access to the fact that the influence of his racist views on his final verdict is (typically) not consciously accessible. The fact that the influence of racism on belief is not (typically) accessible is typically accessible.[13] It is this second-order fact that matters when assessing the justification of Jones' belief, especially, again, when the standards for justification are as high as in Puddifoot's scenario. That the impact of implicit biases on other cognitive states is a consciously *in*accessible factor is a consciously accessible factor. It is a consciously accessible factor, on which Jones fails to reflect, thus forming an unjustified belief.

The move here is to keep the supervenience base for the justification of a belief restricted to what is accessible to the subject in the traditional, narrow fashion—justification supervenes on what is introspectively accessible to the subject—but to understand that what determines whether or not Jones' belief about the black defendant is justified is not the very impact of Jones' racist bias on his belief—a fact that is typically not accessible. What determines whether or not Jones' belief is justified is the fact that such an impact is (typically) inaccessible, and this latter, second-order fact is accessible to Jones. Jones just fails to access it when he could. So Jones' belief is not justified.

This second strategy is inspired by a characterization of accessibilism recently defended by Michael Hatcher (2016). According to Hatcher, the thesis that whether *S* is justified to believe *p* is determined by what is accessible to *S* is ambiguous between the following two readings (Hatcher, 2016, p. 5):

(A)$^{\text{very things}}$ Whether *S* is justified to believe *p* is determined by the very things accessible to *S*.

(A)$^{\text{facts about}}$ Whether *S* is justified to believe *p* is determined by the facts about which things are accessible to *S*.

Hatcher illustrates the general kind of ambiguity that motivates the distinction between (A)$^{\text{very things}}$ and (A)$^{\text{facts about}}$ with a couple of examples. In the first one we are asked to consider a sentence like (a).

[13] Perhaps, it would be more appropriate to say, as an anonymous referee suggests, that what is obviously introspectively accessible is the lack of introspective accessibility of the influence of racist biases on verdicts in a wide range of cases.

(a) Whether Abby is ready for a history exam is determined by what Abby knows.

Although it is possible to interpret (a) as saying that whether Abby is ready to take the exam is determined by the very things Abby knows (thus making the sentence obviously false), it makes more sense to think that whether Abby is ready to take the exam is determined by facts about which things she knows, i.e., the fact that she knows that a certain battle took place at a certain time, that she knows the main contenders in WWII, etc. By contrast, when considering (b), the opposite is the case:

(b) Whether the star will go supernova is determined by what Abby the astrophysicist believes.

Here, the uncharitable reading is to take (b) as saying that whether the star will go supernova is determined by facts about which things Abby believes. Such a reading makes (b) evidently false. On a more charitable reading, (b) says that whether the star will go supernova is determined by the very things Abby believes: some of which are astrophysical facts (Hatcher, 2016, pp. 4-5).

Hatcher (2016) argues in favour of the 'facts about' disambiguation as the correct characterization for accessibilism, as only this reading, he contends, can avoid one of the central objections against the view—the objection that accessibilism involves an infinite regress of facts that must be accessible to the subject. Furthermore, Hatcher argues that just the 'facts about' reading fittingly responds to the main motivations for endorsing accessibilism in the first place. Hatcher's rich argument goes far beyond the scope of this paper. What interests me about his proposed disambiguation is that it allows us to appreciate that accessibility (or inaccessibility) to first-order facts does not necessarily carry over to accessibility (or inaccessibility) to second-order facts and that it makes much more sense to view the accessibilist notion of justification as pertaining to the latter kind of fact, thus keeping accessibility framed in standard introspective terms.

That the very impact of implicit biases is not accessible by reflection alone allows Puddifoot's argument to gain some initial plausibility. But what matters for assessing whether Jones' belief in ACCESS_2 is justified is not that the impact itself is not consciously accessible. It is not the very things that are accessible to Jones that

determine whether he is justified or not. Rather, it is the facts about which things are accessible to him. The very impact of implicit racism on Jones' thinking may not be accessible to Jones, but the inaccessibility to this fact is an accessible fact—an accessible fact that Jones fails to access, thus forming a belief whose justificatory status is not the same as the fame belief in a scenario where there is no intervention of implicit biases. Once accessibilism is thus re-defined, the Jones of this world, in societies like ours, in situations like the one described by Puddifoot, are not justified to believe the beliefs formed as a result of the influence of implicit biases by accessibilism's own lights. Or so it seems.

7. Two objections. Debunking the anti-accessibilist intuition.

Here are two closely related objections one may raise against the two accessibilist-friendly proposals I have just sketched.[14] First, the proposals are too demanding; they both assume an over-intellectualized subject. For the first to work, it will have to be true that ordinary subjects ought to know that implicit biases have a pervasive influence on their thinking even when they are not aware of this influence. The second proposal demands, in the specific case of racism, that people have introspective access to the fact that the influence of their racist views on the verdict they reach is not typically introspectively accessible. In both cases, the strategy seems to work only if we think of highly reflective, socially sensitive, and intellectually sophisticated subjects, perhaps academics, politicians or educators working on implicit biases, but it hardly seems true of ordinary subjects. We have to remember that it is not explicit racism or sexism or homophobia that we are discussing here. Maybe ordinary subjects, at least in contemporary societies, are aware of these worrying phenomena, but the issue is rather whether ordinary subjects, in all kind of societies, are aware of *implicit* biases, their pervasive influence on their thinking and behaviour, and the fact that they are not, for the most part, aware of this influence.

The proposals—the second objection goes—rely just on contingent facts about awareness of the content of implicit biases and public availability of information about how unaware we are of their influence on thought and behaviour. The

[14] I thank two anonymous referees of this journal for pressing this question.

contingency of these facts, however, will not solve the problem raised by Puddifoot style scenarios vis-à-vis accessibilism. For accessibilism, as a fundamental normative thesis is, if true, necessarily true. Yet, the two suggested proposals only dent premise 2' in ACCESS_2 contingently. It is still possible that, for some agents, in some epistemic situations, the impact of implicit biases on other cognitive states and behaviour is a consciously *in*accessible factor. Hence, accessibilism is not necessarily true. Therefore, accessibilism is false. In other words, for Puddifoot's argument to work, she only needs to show that, in some cases, the influence of implicit biases is not accessible. I need to show much more. I need to show that such an influence is accessible in all cases. Yet, my proposals can only guarantee this under certain social conditions and perhaps only for subjects with a certain cognitive sophistication.[15]

These are both important points and I grant them unreservedly. They show that my attempt to argue against premise 2' falls short of showing its falsehood. The suggested accessibilist-friendly strategies only show that it is possible that the impact of implicit biases on other cognitive states is a consciously accessible factor. But to appropriately relate to 3', this is too weak. 3' says that, in all cases, the impact of implicit biases on other cognitive states is relevant to epistemic justification. Hence, the accessibilist is still in trouble. Yet, the strategies reveal something general and important about our epistemic obligations and about the intuitions elicited by the role of implicit biases in accessibilist justification. In order to show that, I now move briefly to a discussion of 3'.

As I said earlier, taken at face value, 3' just begs the question against accessibilism. Yet, the intuition behind Puddifoot style scenarios is meant to force us to accept it. Denying 3', Puddifoot says, would just commit us to giving up one of the main motivations for accessibilism, i.e., deontologism about justification (2016, pp. 426-427). This is how I see the problem. Denying 3' would place the influence of implicit biases on a par with the action of an evil demon, but there is a significant difference between the influence of implicit biases and the influence of (new) evil

[15] I say "perhaps" to acknowledge Hahn et al.'s (2014) suggestion, mentioned in footnote 7, that remarkably naïve subjects are still surprisingly able to predict how their implicit attitudes will influence their behaviour in different experimental settings, even when they do not even seem to have a clear notion of what implicit attitudes are. This acknowledgement, however, still fails to show that the influence of implicit biases is accessible in all cases, regardless of cognitive sophistication.

demons. Or, at least, if we feel there is a significant difference, we do so because we take the former to be a cognitive vice, manifested in our reasoning and evidence evaluation, while the latter is just an unlucky accident that leaves our rational abilities untouched.[16] If we can, at least in principle, eradicate or aspire to eradicate cognitive vices, but we cannot remove the doings of evil demons, then there must be some truth in the idea that implicit biases are relevant to epistemic justification.

It is with regard to this difference that the charge of over-intellectualism to the two accessibilist-friendly strategies defended here teaches us an important lesson. For, although there seems to be some truth in the idea that implicit biases are relevant to epistemic justification, what moves us to feel this way is an intuition based on an over-intellectualized picture of us, as subjects of epistemic obligations. Subjects who do not have the cognitive resources to amend vices or just are not surrounded by the right social structures—either because their social environments do not exhibit the kind of structural social injustice that lies behind most biases or because, even if they do exhibit it, such societies have no mechanisms that could play the role of Gibbons' fridge note—seem to be in the clear, by deontologism's own lights.[17] If we continue to feel the force of the intuition behind the anti-accessibilist premise 3', it is precisely because we, as philosophers, as academics, as sophisticated cognitive agents, are in a position to know about the content and the characteristic inaccessibility of the influence of implicit biases.

The intuition behind 3' is thus fuelled, ultimately, by the same kind of over-intellectualism that affects the recommended accessibilist-friendly strategies. Were we to remove this over-intellectualist overtone, the intuition would have much less of a pull, if it remained at all. Just this would be, of course, good news for the accessibilist. But, what I find most interesting in this dialectics is not whether throwing back and forth the charge of over-intellectualism debunks accessibilist or anti-accessibilist intuitions about epistemic obligations. What I find most interesting

[16] Interestingly, if we do have different intuitions about the two sorts of scenarios, this will seem to suggest that the factors relevant to justification must be internal, even if they need not be accessible. Thank you to Conor McHugh (in conversation) for making this point.

[17] There is a sense in which, if the influence of implicit attitudes were completely unknowable to the subject or the subject's peers, if a subject also had no awareness of any facts about their pervasiveness, denying 3' would be justified, as this scenario would very much be like the new evil demon scenario. Thank you to an anonymous referee for pressing me on this point.

is that, if we grant that accessibilism could meet the challenge of implicit biases in the two forms suggested here, but only for cognitively refined and socially-sensitive minds (or only for cognitively refined and socially-sensitive minds in the right environment), then we will also have to grant that awareness of the relevant facts comes in degrees, for cognitive sophistication and social complexity do come in degrees. And, if so, we will also have to grant that lack of cognitive refinement or appropriate environment entails lack of epistemic obligation. Accessibilism thus delivers the right verdict while holding onto the maxim of no justification (remember it is just propositional justification, not doxastic justification, and not knowledge) without access to the relevant justifiers. If we have good reasons to believe *p* given the available evidence, then we will be justified to believe *p*, even if a potentially relevant justifier, which is accessible (in either of the two forms advocated here) to a more cognitively sophisticated subject, is inaccessible to us. *We* would not have failed to fulfil any epistemic duty on this scenario. But we would have failed to fulfil our epistemic duties, if we, as sophisticated cognitive agents, had access to *all* potentially relevant justifiers, yet failed to take them into account.[18]

[18] This final picture about the gradual nature of awareness, and hence accessibility and justification, fits nicely Madva's (2017) view about the gradual nature of our moral responsibility for implicit biases. It also fits standard moral judgments about e.g. racial discrimination, as shown by some experimental philosophy studies run by Cameron, Payne and Knobe (2010). They found that, when implicit attitudes were characterized as completely unconscious, participants were more inclined to think that people under their influence were less morally responsible than when the influence was taken to be conscious but difficult to control.

References

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie* 48: 145–160.

BonJour, L. 1980. Externalist theories of empirical knowledge. *Midwest Studies in Philosophy* 5: 53–73.

BonJour, L. 1985. *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press.

Brownstein, M. & Madva, A. (2012). Ethical Automaticity. *Philosophy of the Social Sciences* 42(1): 67–97.

Cameron, C. D., Payne, B. K. & Knobe, J. (2010). Do theories of implicit race bias change moral judgments? *Social Justice Research* 23: 272–289.

Carruthers, P. (2017). Implicit versus explicit attitudes: Differing manifestations of the same representational structures? *Review of Philosophy and Psychology*. DOI 10.1007/s13164-017-0354-3

Chisholm, R. 1988. The indispensability of internal justification. *Synthese* 74(3): 285–296.

Cohen, S. (1984). Justification and truth. *Philosophical Studies* 46: 279–295.

Conee, E., and R. Feldman. 2004. Internalism defended. In *Evidentialism: Essays in Epistemology*. New York: Oxford, 53–82.

Cunningham, W. A., Nezlek, J. B., & Banaji, M. R. (2004). Implicit and explicit ethnocentrism: Revisiting the ideologies of prejudice. *Personality and Social Psychology Bulletin,* 30: 1332–1346.

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass* 8(7): 342–353.

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

Feldman, R. 2005. Justification is Internal. In *Contemporary Debates in Epistemology*. eds. Matthias Steup and Ernest Sosa. Malden, MA: Blackwell. pp. 270–284.

Gawronski, B., W. Hofmann, & C. Wilbur (2006). Are "implicit attitudes unconscious? *Consciousness and Cognition*, 15: 485–499.

Gendler, T. (2008a). Alief and belief. *The Journal of Philosophy* 105(10): 634–663.

Gendler, T. (2008b). Alief in action (and reaction). *Mind and Language* 23(5): 552–585.

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

Gibbons, J. (2006). Access externalism. *Mind* 115(457): 19–39.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102*:* 4–27.

Greenwald, A. G., Banaji, M. R. & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology* 108: 553–561.

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

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L. & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology* 97: 17–41.

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.

Hall, D. L. & Payne, B. K. (2010). Unconscious influences of attitudes and challenges to self-control. In R. R. Hassin, K. N. Ochsner & Y. Trope (eds.) *Self Control in Society, Mind, and Brain* (pp. 221–242). Oxford: OUP.

Hatcher, M. (2016). Accessibilism defined. *Episteme* doi:10.1017/epi.2016.36

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measure. *Personality and Social Psychology Bulletin* 31: 1369–1385.

Hughes, S., Barnes-Holmes, D. & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record* 61(3): 465–498.

Kassin, S., Fein, S., & Markus, H. R. (2010). *Social Psychology* (8th ed.). Belmont, CA: Wadsworth Cengage Learning.

Kenrick, D. T., Neuberg, S. L., & Cialdini, R. B. (2010). *Social psychology: Goals in interaction* (5th ed.). Boston, MA: Allyn & Bacon.

King, M. & Carruthers, P. (2012). Moral responsibility and consciousness. *Journal of Moral Philosophy*, 9(2): 200–228.

Lehrer, K. & Cohen, S. (1983). Justification, truth and coherence. *Synthese* 55(2): 191–207.

Levy, N. (2014). Neither fish nor fowl: implicit attitudes as patchy endorsements. *Noûs* 49(4): 800–823.

Madva, A. (2016). Why implicit attitudes are (probably) not beliefs. *Synthese* 193: 2659–2684.

Madva, A. (2017). Implicit bias, moods, and moral responsibility. *Pacific Philosophical Quarterly*. DOI: 10.1111/papq.12212

Mandelbaum, E. (2016). Attitude, association, and inference: On the propositional structure of implicit bias. *Noûs* 50(3): 629–658.

McConnell, A.R., Dunn, E.W., Austin, S.N., & Rawn, C.D. (2011). Blind spots in the search for happiness: Implicit attitudes and nonverbal leakage predict affective forecasting errors. *Journal of Experimental Social Psychology*, 47(3): 628–634.

Mitchell, C., De Houwer, J. & Lovibond, P. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences* 32(2): 183–198.

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes and Intergroup Relations* 8: 39–52.

Nosek, B. & M. Banaji, 2001, The go/no-go association task. *Social Cognition*, 19(6): 625–666.

Oswald, F., Mitchell, G., Blanton, H., Jaccard, J. & Tetlock, P. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology Studies* 105(2): 171–192.

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.

Payne, B., C.M. Cheng, O. Govorun, & B. Stewart (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89: 277–293.

Payne, B. K., Jacoby, L. L., & Lambert, A. J. (2004). Memory monitoring and the control of stereotype distortion. Journal of Experimental Social Psychology, 40: 52–64.

Pryor, J. (2001). Highlights of recent epistemology. *British Journal for the Philosophy of Science* 52: 95–124.

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

Quillian, L. (2008). Does unconscious racism exist? *Social Psychology Quarterly*, 71(1): 6–11.

Ranganath, K., Smith, C., & Nosek, B. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology* 44: 386–396.

Smith, A. (2012). Attributability, answerability, and accountability: In defense of a unified account. *Ethics* 122(3): 575–589.

Steup, M. (1996). *An introduction to contemporary epistemology*. Upper Saddle River, NJ: Prentice-Hall.

Steup, M. (1999). A defense of internalism. In *The Theory of Knowledge: Classical and Contemporary Readings*, 2nd ed. Belmont, CA: Wadsworth, 373–84.

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