
Article:
Holroyd, J., Scaife, R. and Stafford, T. (2017) What is implicit bias? Philosophy Compass, 12 (10). e12437.

https://doi.org/10.1111/phc3.12437

What is Implicit Bias?

Journal:	*Philosophy Compass*
Manuscript ID	PHCO-1075.R1
Wiley - Manuscript type:	Article
Keywords:	Naturalistic Philosophy < Compass Sections, Mind and Cognitive Science < Philosophy < Subject, Cognitive Science < Mind and Cognitive Science < Philosophy < Subject, Philosophy of Mind < Mind and Cognitive Science < Philosophy < Subject, Naturalistic Philosophy < Philosophy < Subject
Abstract:	Research programs in empirical psychology over the past few decades have led scholars to posit implicit biases. This is due to the development of innovative behavioural measures, which have revealed aspects of our cognitions that may not be identified on self-report measures requiring individuals to reflect on and report their attitudes and beliefs. But what does it mean to characterise such biases as implicit? Can we satisfactorily identify the grounds for identifying them as bias? And crucially, what sorts of cognitions are in fact being measured; what mental states or processes underpin such behavioural responses? In this paper, we outline some of the philosophical and empirical issues engaged when attempting to address these three questions. Our aim is to provide a constructive taxonomy of the issues, and how they interrelate. As we will see, any view about what implicit bias is may depend on a range of prior theoretical choices.

What is Implicit Bias?[1]

1. Introduction

Research programs in empirical psychology over the past few decades have lead scholars to posit *implicit biases*. This is due to the development of innovative behavioural measures, that have revealed aspects of our cognitions which may not be identified on self-report measures requiring individuals to reflect on and report their attitudes and beliefs. But what does it mean to characterise such biases as *implicit*? Can we satisfactorily identify the grounds for identifying them as *bias*? And crucially, what sorts of cognitions are in fact being measured; what mental states or processes underpin such behavioural responses?

In this paper, we outline some of the philosophical and empirical issues engaged when attempting to address these three questions. Our aim is to provide a constructive taxonomy of the issues, and how they interrelate. As we will see, any view about what implicit bias is may depend on a range of prior theoretical choices. First, let us get some paradigm cases of the phenomena at issue on the table.

2. The phenomenaon

These are the sorts of behavioural indirect measures that have provided evidence of the existence of implicit biases,[2] and serve t reveal can stand in as paradigms of the sort of psychological phenomena that philosophers have engaged with.

a. Implicit Association Tests

Perhaps the most well-known of these measures, the Implicit Association Test (IAT) has been participated in millions of times, both in laboratory studies and via online testing hubs (such as that run by Project Implicit https://implicit.harvard.edu/implicit/). These studies are essentially categorisation tasks where participants are instructed to classify a target stimulus (terms or images) into one of two pairs of categories. For example a race IAT asks participants to press one key to classify a target stimulus as belonging to the disjunctive category 'white or negative' and a different key to classify the target as 'black or positive'. Then the category pairings switch: one key represents the disjunctive category 'white or positive' and the other 'black or negative'. Participants are instructed to categorise as quickly as possible whilst making as few errors as possible: they face a speed/accuracy trade-off. Responses which are too fast or too slow are eliminated to prevent random fast clicking or any attempt to 'game or manipulate' the test. The speed of categorisation and number of errors made with the first set of disjunctive categories (white/negative; black/positive) are compared with that from the second (white/positive; black/negative).[3] If an individual who is slower and/or makes more errors when black is categorised with positive and white with

[1] This paper was produced as part of a Leverhulme Trust research project grant on Bias and Blame (RPG-2013-326). We are grateful to the Leverhulme Trust for their support. Acknowledgements removed.

[2] Note that this is not the only sort of evidence available to us: testimonial evidence, from victims and witnesses, about unintentional or unwitting discrimination by people who professed non-discriminatory attitudes is pervasive and predates the upsurge of attention from empirical psychologists. See Holroyd & Puddifoot (forthcoming) for discussion of the problems attendant upon the way philosophers have treated these different sources of evidence.

[3] For details of the IAT scoring algorithm see Greenwald, A. G. et al (2003).

negative, than when white is categorised with positive and black with negative, it is inferred from this pattern of responses that she has more accessible, and therefore other things being equal stronger, unconscious associations between black people and negative evaluations.[4]

The majority of people racialised white perform in this way, which suggests that they have stronger negative associations with black people.[5] In participants racialised black, approximately half have negative associations with black people.[6] Moreover, in black participants, the overall pattern of responses is interpreted as a demonstrating weak rather than strong preference for white over black people (Nosek, Banaji & Greenwald, 2002: 105). These patterns of response have~~is has~~ been consistently found both when race is represented by images (pictures of black and white faces) and when it is represented lexically (stereotypically black- or white-sounding names).[7]

IATs and other indirect measures are used to access not only associations with stigmatised social groups (such as gender, ethnicity, age, disability), but also associations between various target concepts. The Project Implicit site includes flower/insect IATs, for example; and such measures have also been used (e.g.) to better understand brand associations by the marketing industry. Other indirect measures include semantic priming measures (Banaji & Hardin, 1996), the Affect Misattribution Procedure (Payne et al 2005), and the Go/No-Go Association Test (Nosek & Banaji 2001).[8]

b. hiring decisions

~~So~~Indirect measures such as the IAT are taken to reveal implicit attitudes, which are hypothesised to underpin discriminatory behaviours.~~me studies have monitored behaviours in contexts intended to more closely approximate the 'real world'.~~ Some studies focus on the behaviours that might be influenced by implicit biases. For example, in one study individuals were asked to evaluate the qualifications and credentials of potential job applicants, and report back on how likely they were to recommend that the individual be hired. When the applicant's materials indicated they were racialised black, the evaluations were less positive, and fewer hiring recommendations made, than when the applicant was indicated as racialised white (Dovidio & Gartner 2000. See also Bertrand & Mullainathan, 2004). Steinpreis et al (1999) sent out CVs (for an early career researcher) to academic psychologists in the US. The CVs were identical, but sometimes identified the applicant as female, other times as male. Those rating the CVs were more likely to evaluate positively the very same CV when it had a male name at the top (see also Moss-Racusin et al 2012). In

[4] For recent discussion of the predictive validity of the IAT, and its significance, see the roundtable discussion at The Brains Blog: http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx

[5] Note that some authors have distinguished between 'associations' (taken to be connections between concepts, such as a social group and stereotypical content) and 'evaluations' (taken to be connections between positive or negative notions and the target object). Again, see the FAQ at Project Implicit for an example of this framing. https://implicit.harvard.edu/implicit/demo/background/faqs.html#faq6 See also Amodio & Devine 2006. However, it is unclear whether this characterisation is theoretically viable, and much will depend on the view one takes of what is measured by the IAT. See Holroyd & Sweetman 2016, and Madva & Brownstein 2016, for discussion of whether such ways of categorising the states measured are defensible.

[6] See the discussion of responses gathered from white and black participants at Project Implicit: https://implicit.harvard.edu/implicit/demo/background/faqs.html#faq19

[7] Dovidio et al, (1997); Nosek,et al.~~.~~ (2007).

[8] For a useful discussion of these measures, see: Brownstein 2017

~~both~~such studies, these effects were found in cases where there was room for discretion - the applicant was neither obviously 'stellar' nor obviously unqualified. And, in both cases these effects were found notwithstanding the participants' respective self-reported anti-racist attitudes, or commitments to evaluating objectively the applicants. If these self-evaluations are to be taken at face value, it would appear that the participants do not realise that they are under-evaluating black or female candidates, nor do they intend to do so. The effects have been attributed, then, to implicit biases, rather than explicit prejudice.[9]

c. Microbehaviours

A cluster of studies have examined responses that involve affective processes and their manifestation in behavioural responses. These are studies on 'micro-behaviours'. They have tended to focus on ~~–~~ unintentional non-verbal behaviours that manifest tension or discomfort.[10] For example, the eye-blink rate of an individual, the extent to which she engages in fidgeting behaviours, and how closely she positions herself to other individuals, are instances of these non-verbal behaviours. Such behaviours are often automatic, and not under the intentional guidance of the agent. White people have been found to display these micro-behaviours to greater degrees in interracial interactions with black interlocutors, affecting the quality of those interactions (Dovidio, Kawakami & Gartner 2002; Dovidio, Gaertner, Kawakami & Hodson, 2002).

Whilst these studies help us to get a handle on the phenomena at issue, there remain the questions to which we seek precise answers: first, in what sense are the mental goings on which produce these responses *implicit*? Second, what are the criteria we should use to diagnose a response as *biased*? And finally, how should we characterise the psychological phenomena underpinning these responses? In the following sections, we address these questions, identifying the relationships between the different answers.

3. Theoretical choices

It is worth spelling out different aims that appear to inform the various answers that have been given to these questions. These have rarely been made explicit, but rather appear to be tacit or unarticulated assumptions that theorists have in mind as desiderata for a successful account of implicit bias. We state them here, but go on to tease out their relationship to the accounts presented below:

D1: to distinguish implicit from explicit mental states or processes.

D2: to capture interesting cases of dissonance between agent~~'~~s' professed values and the cognitions driving responses to these measures.

D3: to formulate interventions for changing bias, or blocking discriminatory outcomes.

[9] For a helpful overview of a range of behavioural effects taken to be influenced by implicit biases, see Jost et al 2009.

[10] Whilst the studies we mention have focused on these sorts of micro-behaviour, note that micro-behaviours could also manifest positive attitudes, and could include verbal behaviours (such as slips of the tongue). Thanks to an anonymous reviewer for emphasising this point.

D4: to accommodate or explain the full range of the phenomena captured by indirect measures.

D5: to gain traction in addressing problems of marginalisation and under-representation, and draw attention to complicity in these problems.[11]

In broader terms, it is worth noting that some of these aims are directed at capturing some unified psychological kind assumed to be operative in the phenomena described (D1-D3); others are more focused on pragmatic or political aims (D4, D5). We grant legitimacy to each of these aims: but as we will see, these desiderata cannot all be met at once, so theoretical choices have to be made about what one wants an account of implicit bias to do. These choices govern the appeal of certain ways of characterising *implicit, bias* and the psychological reality that these states or processes may have.

4. What is *implicit?*

What does it mean to identify some state or process as implicit? In this section, we aim to show three things. First, that that there are various competing ways of characterising the implicit; second, that the choice one makes about how to characterise the implicit depends on prior theoretical choices about the phenomena one is aiming to capture. Finally, no one view unproblematically carves our cognitions into implicit and explicit.

a. Implicit as unconscious

~~The most~~A common understanding of the *implicit*, both amongst lay-persons, psychologists, and philosophers, is equivalent to *unconscious*. To see how this understanding has informed public discourse and lay-understandings of the phenomena, observe that Wikipedia defines implicit attitudes as "evaluations that occur without conscious awareness...".[12] Scholarly but public-facing sources such as ~~T~~the frequently asked questions page of the Project Implicit website explains that implicit attitudes are: "positive and negative evaluations that occur outside of our conscious awareness and control".[13] In academic publications, Gawronski et al, noted that "a widespread assumption underlying the application of indirect measures is that they provide access to unconscious mental associations that are difficult to assess with standard self-report measures" (2006, p486).[14] Many philosophers also use implicit in this sense. For example, Machery et al describe implicit biases as "biases [individuals] are not aware of having" (2010, p.227) and Kelly characterises implicit biases as "outside of person's conscious awareness" (2013, p.460). The initial appeal of this characterisation is that it seems to provide a helpful way of distinguishing implicit from explicit - consciously held - states or processes. This characterisation may be driven by desiderata:

[11] This list is not intended to be exhaustive, but we identified these aims as the most salient - albeit usually unstated - to the accounts considered below.

[12] https://en.wikipedia.org/wiki/Implicit_attitude

[13] https://implicit.harvard.edu/implicit/faqs.html#faq0

[14] The authors Gawronski cites as making this assumption include: Bacchus et al., 2004, Banaji, 2001, Bosson et al., 2000, Brunstein and Schmitt, 2004, Cunningham et al., 2004, Greenwald and Banaji, 1995, Jost et al., 2002, Phelps et al., 2000, Rudman et al., 1999, Spalding and Hardin, 1999, Teachman et al., 2001 and Wilson, 2002

D1: to distinguish implicit from explicit mental states or processes.

Moreover, seeing implicit biases as unconscious explains the fact that many people find the outcomes of indirect measures surprising, given their failure to accord with their professed values. This understanding is further motivated by the desiderata:

D2: to capture interesting cases of dissonance between agent's professed values and the cognitions driving responses to these measures.

However, note that for some theorists (including some mentioned above) the implicit is not *necessarily* unconscious; rather it is characterised as *typically* outside of conscious awareness. For example, Brownstein and Saul introduce implicit bias as "a term of art referring to evaluations of social groups that are largely outside of conscious awareness or control" (2016, p1). Whilst characterisations based on what is typical are not vulnerable to counterexamples in the same way as those which see bias as necessarily unconscious, they thereby offer a less principled basis for drawing the implicit-explicit distinction.

A further concern with understanding *implicit* as *unconscious* is that there are various possibilities regarding *what* the agent is unconscious of (see Gawronskia et al, 2006; and Holroyd, 2014). Contenders include that the individual lacks consciousness of:

i. The bias itself (the mental state, process or trait)
ii. The influence of the bias on the decision or action (whether or how the decision or action has been influenced)
iii. The source or cause of the bias

Regarding i. evidence suggests that individuals may be able to accurately predict their own IAT outcomes, without having any previous experience of taking an IAT on which to base their prediction (Hahn et al 2013). This suggests individuals have some awareness of the cognitions revealed on such measures. Furthermore, if lacking consciousness in this sense is a requirement, rather than typically true, of implicit biases, this entails that a bias is no longer implicit once a person has becomes aware that they have it.[15]

Regarding ii. Hahn et al (2013), Monteith et al (2001) and Scaife et al (ms.~~forthcoming~~) both found that participants self-reports of their experience of taking the IAT accurately tracked their IAT scores. In the latter study a number of participants reported experiencing difficulties in the incongruent (counter-stereotype) blocks of the IAT. Since in such studies, participants were aware of experiencing incongruent blocks of the IAT as more difficult, this suggests that individuals can gain awareness of the influence that biases exert on their responses, and can accurately report on this influence. On these occasions, individuals are not unaware of the influence of the bias on their responses~~, suggesting that the influence of biases is not necessarily outside conscious awareness.~~.

[15] Another contender for the sense of awareness relevant is that individuals lack awareness of the body of research on implicit bias. Similarly, on this characterisation, biases will ~~seek~~cease to be implicit once people learn of this research.

Regarding iii. it is likely that few people have awareness of how their cognitions came to be structured in the ways revealed by indirect measures. However, this will not distinguish implicit from explicit cognition, since few people are aware of the source of their explicit mental states and decisions (see Wilson 2002; Johansson et al, 2005; Carruthers 2010 & 2011; and Hall et al, 2012). So if one is motivated by D1, this characterisation will lose its appeal.

In sum: various available understandings of the *implicit* as *unconscious* fail to cohere with empirical evidence about the sort of awareness individuals appear to have; or fail to pick out a feature that distinguishes it from the explicit states (those typically thought of as conscious).

b. Implicit as beyond control

An alternative understanding of *implicit* pertains to the kind of control that individuals have - or lack - with respect to the states or processes at issue. Some have suggested that what characterises states and processes as implicit is that they are beyond direct control: an individual cannot remove or prevent the impact of an implicit bias through an act of will (i.e. simply by choosing to do so) (Saul 2013, Kelly & Roedder 2008).[16] The claim is not that individuals have no control at all over implicit biases; a number of authors have noted that implicit biases are still subject to various forms of indirect control (see Faucher 2016, Holroyd 2012, Brownstein 2015) and ecological control (see Holroyd & Kelly, 2016~~5~~).

Characterising the implicit in these terms seems to be motivated both by a concern to contrast with the explicit (D1), and to explain dissonance (D2) - because those actions influenced by implicit biases cannot be guided by the agent's reflective values. Note also that the focus on kind of control is also useful for meeting a further aim:

> D3: to formulate interventions for changing/mitigating bias, or blocking discriminatory outcomes.

If it is definitive of biases that they are beyond our direct control, this firmly focuses attention on new kinds of indirect strategies that are more likely to succeed in mitigating the role of bias, or insulating outcomes from its influence.[17]

Whether such a characterisation can meet these desiderata (in particular, D1 & D2), depends on whether we can in fact cleanly distinguish between direct and indirect control. Moreover, even if this distinction can be satisfactorily made, it is not clear that it would cut in an intuitive place between the implicit and explicit. Many other cognitive processes and states which one might expect to fall on the explicit side of the distinction are also beyond our direct control. Whilst some beliefs can be changed at will, others cannot - yet the inclination to describe beliefs as explicit or implicit does not appear to correspond to this feature, namely, whether and how they can be altered. Characterising the implicit as beyond

[16] In fact, there is evidence indicating that trying to suppress implicit bias through an act of will can have a rebound effect (Follenfant & Ric 2010).

[17] Other considerations may further motivate the move to characterise implicit biases in terms of control rather than unconsciousness: see Duguid and Thomas-Hunt (2015) for concerns that merely raising awareness about bias can in fact worsen the problem. In any case, as Saul rightly remarks: "even once [individuals] become aware that they are likely to have implicit biases, they do not instantly become able to control their biases…". (2013a, 55)

direct control, then, appears to be over-inclusive. ~~Many - perhaps all - beliefs and attitudes cannot be immediately controlled by acts of will.~~

However, if one is motivated primarily by D3, and gives little weight to D1 and D2, then it may not matter if the criterion is over--inclusive, and extends also to include some states or processes typically thought of as explicit.

c. Implicit as dissonant/unendorsed

One might identify the implicit as being *dissonant* with the agent's other (or central) cognitions; implicit biases as unendorsed. Glasgow emphasizes that people are *alienated* from their biases in that "they sincerely and truly claim that their biases do not represent who they *really* are" (2016, p37). Similarly, Frankish has proposed that we should understand biases as implicit when an agent "does not endorse it in her conscious reasoning and decision making" (2016, p.25). And Levy argues that the type of processing characteristic of implicit biases prohibits them from truly reflecting the agent, because their associative structure excludes them from being subject to 'rule based-processing'. Accordingly, for Levy, biases are 'patchy' and cannot be inferentially sensitive to, or integrated with, the agent's other evidence-sensitive attitudes (2015, p.812-816).

It is clear enough that seeing dissonance as characteristic of implicit biases is primarily motivated by the aim:

> D2: to capture interesting cases of dissonance between agent~~'~~s' professed values and the cognitions driving responses to these measures.

This difference between behavioural responses on indirect measures, and self-reported attitudes draws a lot of attention; it is perhaps the most striking feature of paradigm cases of implicit bias. Furthermore, there are good pragmatic reasons to focus on such cases because those who disavow their biases are likely to be highly motivated to take whatever steps are necessary to prevent their implicit biases influencing them.

But it should also be clear that one cannot at the same time meet this aim whilst also meeting some of the other desiderata. In particular, accounts that characterise implicit biases in terms of dissonance fail:

> D4: to accommodate or explain the full range of the phenomena captured by indirect measures.

Implicit biases may align or in accord with explicit attitudes (Zheng 2016; Holroyd 2016). While such cases have received little attention there are a number of studies which indicate that such alignment is to be expected: Devine et al (2002) found that implicit race biases were stronger in individuals who showed more explicit racial prejudice on self-report measures. Yet there will be no implicit bias to speak of in these cases, if the implicit is characterised in terms of dissonance. For example, in considering cases in which implicit biases become integrated with the agent's explicit attitudes, becoming 'annexed' to her endorsed attitudes, Levy writes that 'it is an open question whether implicit attitudes survive such annexation: an annexed attitude possesses the appropriate set of inferential relations to other attitudes, and thereby cease to be a patchy endorsement. If implicit attitudes are always patchy endorsements, such an annexed attitude transforms into an ordinary attitude

(conscious or not)' (2017, 21. We discuss Levy's notion of patchy endorsements below). Hence we can see that whilst reserving the notion of implicit bias for cases in which there is dissonance, or a failure of 'annexation', is legitimate, the resulting characterisation will be ill-placed to also meet D4.[18]

d. Implicit as accessed by certain kinds of measure.

Given these difficulties, we might simply think that the domain of the implicit is delineated by what is revealed on indirect measures, such as those earlier described. This strategy is deployed by some psychologists~~., who avoid giving a precise definition of what makes biases *implicit*.~~ For example, Fazio and Olson claim that "…it is more appropriate to view the measure as implicit or explicit ... What makes priming or the IAT implicit is that these techniques provide estimates of individuals' attitudes without our having to directly ask them for such information" (2003 p.303). Note, however, that Fazio endorses a view whereby the measures alone are implicit or explicit; the representations measured should not themselves be conceived of in those terms (Fazio & Towles-Schwen 1999). If one seeks to draw *that* distinction (motivated by D1) one could ~~We could~~ say (departing from Fazio's usage)~~, then,~~ that certain~~the~~ measures are implicit, and whatever is revealed by those measures then inherits the label 'implicit'. Alternatively, we might say a bias is implicit if it is revealed by indirect measure.[19] Such a view is appealing insofar as one is motivated by

D4: to accommodate the full range of the phenomena captured by indirect measures.

Such a characterisation will *a fortiori* capture all the results of indirect measures as falling within the domain of implicit cognition. This is attractive because, as we have seen, it is tricky to specify a principled way of defining the implicit.

Note, though, that in understanding *implicit* in this sense - as that detected by indirect measures - we gain virtually no theoretical insight into the properties that such states may possess, nor how or whether they are distinct from states typically thought of as explicit. This is because the fact that they can be accessed by an indirect measure does not tell us that they can *only* be accessed in this way. That they are revealed in automatic responses does not tell us that they cannot be controlled; that they can operate without the agent's awareness does not tell us that they always do. Accordingly this characterisation does not help us to delineate the properties of implicit biases, nor whether they are such that other measures (direct or self-report measures) cannot access them. Moreover, such a view provides no conceptualisation of what such states are nor when and why they might diverge from other attitudes. In particular, no information is provided about why we might need, or at least be better able, to access them via indirect measure rather than some other means. This hollowed out notion of the implicit therefore provides no insight into when we might doubt the accuracy of self-report measures, nor why, on those occasions, we might expect implicit measures to provide different, or more (or less) predictive results from those garnered through self-report.

[18] See Holroyd 2016 for discussion of this aspect of Levy's view, and in particular of whether it is able to make fine-grained distinctions, in the moral assessments of agents who harbour implicit biases, that we might hope to.

[19] Note that Fazio and Olson hold back from these labels, instead suggesting that the terms implicit or explicit should perhaps not be applied to the states or processes, but to the measures alone. On their view, these different measures access the same psychological construct: the agent's attitude.

It may be that in fact, some unstated assumptions are at work when this usage of implicit is endorsed: that implicit measures are those which bypass conscious awareness, or accesstap states that are beyond direct control. But this is to fall back on the understandings of *implicit* that we have problematised above.

e. Implicit as discursively useful

It is worth noting another recent trend in the usage of the term 'implicit', which suggests a rather different agenda from those so far identified (in D1-D4). This usage has most prominently been observed in Hillary Clinton's remarks that 'We all have implicit biases ... What we need to do is be more honest about that, and surface them. Because today, most people believe that they don't have those biases'.[20] Clinton may have had in mind a particular psychological phenomenon when making these claims. But more likely she is using the notion of implicit bias in a way consistent with how the usage of many activists - within and outside of philosophy - and practitioners running training sessions: as a way of opening up discussions about exclusion and marginalisation. Using 'implicit' in this sense usually carries the following important implications: of acknowledging complicity, and taking responsibility. By acknowledging that bias is pervasive, and that everyone may be complicit in perpetuating discrimination and marginalisation, discussion of the problems can avoid finger-pointing or labelling some 'bad' individuals as 'the problem'. This seems to have the pragmatic effect of collectively taking responsibility for the fact that things are not as they should be, and that everyone should have a role in fixing this. The main aim of using *implicit* in this sense, then, appears to be:

> D5: to gain traction in addressing problems of marginalisation and under-representation, and draw attention to complicity in these problems.

Note that this primarily pragmatic, discursively useful notion of the implicit need not take a stand on whether there is any distinctive psychological reality underpinning the responses recorded in indirect measures, that can be distinguished from explicit states or characterised by a unique set of properties. However, using the notion of *implicit* for these pragmatic aims will be hostage to empirical fortune in one sense: it must be true that it is a helpful way of addressing problems of exclusion and marginalisation, and not a distraction from alternative ways of addressing these problems (see Haslanger 2015 for worries along these lines).

In this section, we have argued that there are various ways of characterising the notion of *implicit*, and each is motivated by a somewhat different set of concerns. However, none of these ways of characterising the implicit is wholly satisfactory: which costs or gains one is willing to take on, then, may depend on the aims to which one gives priority.

5. What is *bias?*

[20] http://edition.cnn.com/2016/04/20/politics/hillary-clinton-race-implicit-biases/index.html
See also her remarks in the first presidential debate that 'implicit bias is a problem for everyone': http://www.cnsnews.com/news/article/melanie-hunter/

In this section, we delineate the choices to be made about how to characterise the notion of *bias*, and tease out the ways one's answer to this question is located in relation to the other theoretical choices available.

i. Bias is bad

Many of the scholars working on implicit bias have used the term 'bias' in a way that presupposes that it is something normatively bad. This usage chimes with common-sense understandings of what it is to be biased: in particular, biased *against* someone or some group. For example, Saul writes that 'in the case of women in philosophy, implicit biases will be unconscious biases that affect the way we perceive (for instance) the quality of a woman's work, leading us to evaluate it more negatively than it deserves' (2013a: 40). Biases, here are linked to distorted and negative evaluations.

In another paper, Saul writes that biases 'are unconscious, automatic tendencies to associate certain traits with members of particular social groups, in ways that lead to some very disturbing errors' (2013b: 244). Here, 'bias' is used to capture cases in which there are 'disturbing errors'. In both cases, Saul clarifies that there is another usage of 'bias' whereby bias is not necessarily bad. However, given that the focus is on a certain range of phenomena - those in which stigmatising biases are operative - the theoretical choice is made to use 'bias' to denote cases where something has gone wrong. The task, then, is to spell out exactly what standard implicit biases, in the bad sense, lead us to fall short of. Saul's remarks helpfully suggest two sorts of normative standards: biases might lead us into moral error (when we evaluate people in undeserved ways) or rational error (when we reach false or unwarranted judgements).

a. *Bias* as irrationality

We might want to diagnose bias as bad because it leads agents who are influenced by it into failures of rationality. One might claim that these states or processes constitute a bias since they lead to failures of practical rationality, or the thwarting of an individual's goals: she has the goal of hiring the best candidate, but fails to do so because gender bias inflects her evaluation of the applicant's CV. Note, though that this strategy makes our diagnosis of a state as bias dependent on the agent's goals and values. This captures bias in cases where implicit cognition is dissonant with the agent's values (cf. D2 above); but it may be inapt if one also wants to characterise as bias also those cases in which the bias resonates with the agent's values (see D4 above).

An alternative criteria for identifying the states as 'bias' is to see them as constituting a failure of theoretical rationality, or violation of good knowledge-seeking practice. For example, Saul writes that we are simply making errors because our judgements are 'being influenced by factors [that are] totally irrelevant', namely, social category information (Saul 2013b, 247). Saul spells out this failure in terms of the irrelevance of the factors that are influencing judgement, but there may be various ways of unpacking the failure at issue. We might identify bias in the failure to reliably track the truth; in failures of sensitivity to evidence or of appropriate trust in testimony; failures of epistemic responsibility or to exercise epistemic virtues (see Holroyd & Puddifoot, forthcoming, for articulation of the ways implicit biases might violate a variety of epistemic norms). The point is that there are various ways in which these states or processes might violate norms of inquiry and knowledge seeking: any

of these standards would identify 'bias' in a wider set of cases than the standards of practical rationality.

This position requires defence against the claim that implicit biases, at least sometimes, present no violation of such standards. Rather, they may~~, but rather~~ - insofar as they represent associations learned from our environment - present useful base-rate information about social groups. This is what Gendler suggests, ~~the view that Gendler has proposed,~~ arguing that to try to prevent bias from influencing judgement is, on some occasions, ~~to therefore~~to face ~~some~~ epistemic costs (Gendler 2011). However, as Kelly & Roedder suggest:

> such associations almost always extend beyond what is rational, and there will almost always be a 'remainder': an implicit association that goes beyond what rationality endorses (2008: 530).

What this 'remainder' consists in requires explication, but there are various options available. One may doubt ~~, but one may appeal to doubts about~~ whether implicit biases can encode such statistical data (see Puddifoot ms. for concerns about whether crudely associative states provide such base-rate information). Or one may doubt that~~; or whether~~ they have a structure that renders them appropriately reasons-responsive (see Levy, 2015~~ discussed below~~) and subject to revision in light of evidence (Madva 2016). Thus, the feasibility of one's views about whether a state constitutes a bias - whether it violates norms of rationality or good inquiry - hinges on further questions about what these states are and how they behave.

b. Bias as immoral

In some cases, we might want to identify the cognitions as *bias*, but find that they are less easily diagnosed as violating standards of theoretical rationality, since they are not obviously or always engaged in knowledge seeking contexts: consider the micro-behaviours outlined above. Such behaviours might hinder inquiry (e.g. if they affect interactions involving testimony (cf. Dotson 2012)) but need not; they might instead make for chilly or hostile environments. This might undermine something the agent wants (e.g. an inclusive and respectful workplace). But if one has concerns about spelling out the notion of bias in terms of failures of practical rationality, one might instead see such cases as 'bias' simply because they involve falling short of some moral standard. Kelly & Roedder observe that implicit biases are obviously morally problematic when they lead to harmful or unfair consequences (2008: 527). One reason for focusing on these particular sorts of cognitions and the behaviours they underpin is precisely because of their relationship to patterns of marginalisation, exclusion, and their implication in unjust social structures (cf Saul 2013a; Haslanger 2016). Alternatively, one might see certain states as bias because they are premised on malevolence (Garcia, 2004) or disrespect (Blum, 2004) towards the groups they target (see Kelly & Roedder 2008 for discussion).

Note that articulating such a standard, and seeing failures to meet it as definitive of states that are *biases* again captures a broader set of phenomenon than is included by a standard indexed to the agent's goals (cf. D4).

ii. Bias itself is normatively neutral.

We observed that those who focus on the badness of bias nonetheless acknowledge that *bias* may be used in a broader sense, to capture a wide range of phenomena that may be involved in implicit cognition. For example, in a footnote, Saul observes that one might use the term simply to pick out implicit associations (2013b: 40, fn4). This would be to use the term in a normatively neutral way: to denote a broad set of cognitive phenomena which includes those states or processes that are good or bad: an automatic association between 'salt' and 'pepper' would, on this view, also be an implicit bias. This usage resonates with the idea that the cognitive phenomena at issue extends beyond those involved in social cognition about stigmatised groups. For example, psychologists working on market research have focused on the role of implicit associations in brand preferences and consumer choice (Gregg & Klymowsky 2013). Others have examined the role of implicit associations in the context of health behaviours and policy decisions (e.g. Macy et al 2013, Stacy et al 2000). Indeed, some biases, in this neutral sense, may be indispensable in navigating and understanding the world. The task then is to identify why those which *are* bad are identified as such.

One theorist who proposes such a conception of bias is Antony, who argues that 'bias plays an essential and constructive role in the development of human knowledge' (2016: 158; see also her 1993). On this understanding 'bias' simply means 'a tendency: an inclination of temperament or outlook' (2016: 162) - such tendencies, she argues, are inevitable for limited cognitive agents such as ourselves, and moreover, are often useful in focusing our enquiry on salient possibilities. The key task, then, is to identify which inclinations are innocuous or positively helpful, and which are problematic. Antony's main focus is on which biases incline us away from the truth, and which towards it - and this, she proposes should be uncovered by naturalistic methods: observations of how enquiry proceeds.

In section 3, we argued that the line one takes on what the *implicit* is depends on what aims one has in theorising - what desiderata one is trying to meet. Note that the same is true with respect to which view of *bias* - as bad, or as normatively neutral - one endorses. For example, if one is focused primarily on drawing attention to problematic phenomena (D5), then the usage of bias in the narrower sense (bias as bad) might be efficacious. On the other hand, it may be politically helpful to be able to point out the continuum between implicit biases of the problematic sort and cognitive phenomena on the other (what works will essentially be an empirical matter, and it may depend on the context). Or one might be motivated by wider theoretical aims: for example, Antony is motivated by commitment to a model of enquiry that does not rest on implausible, and unachievable, ideals of 'objectivity', but better descriptively captures how enquiry proceeds. Moreover, whether one sees bias as bad or normatively neutral may depend on, or in turn inform, the view one endorses regarding what psychological reality these states or processes have. We turn to this issue in the final section.

6. What psychological reality might implicit bias have?

Much of the philosophical literature has focused on how to characterise the psychological reality underlying the responses, judgements and behaviours described in the first section. Here, we survey some of this literature, and tease out which of the theoretical and practical choices, identified earlier, appear to underpin these views on the psychological reality of

implicit bias.[21] These choices have often not been manifest in the articulation of these views: our hope is that in elucidating these issues, it makes clearer the commitments taken on by any one such account.

a. beliefs

i. Unconscious beliefs

Some authors have argued that implicit biases are best modelled as familiar mental states: beliefs. For example, Mandlebaum proposes that we consider implicit biases as 'honest-to-god propositionally structured mental representations that we bear the belief relation to' (2015, p.7). However, such beliefs are unconscious. This means that (not necessarily, but at least typically) they do not figure in our conscious thought. Yet, because they are propositionally structured, implicit biases can function inferentially, and in reasons-responsive ways - even whilst beyond the reach of our reflective awareness.

We are now well positioned to identify the theoretical choices that underpin such an account. First, Mandlebaum is clear that one of the key desiderata he is guided by is that of cohering with, and accommodating the empirical evidence (D4). One of the key arguments for his view is that it better explains empirical studies in which implicit biases seem to be operational in inferential reasoning processes. Second, that this view clearly takes a stance on the sense in which these biases are *implicit* and hence differ from explicit thought (compare D1): namely, they are *unconscious*. Finally, note that this view arguably ends up with commitments to the scope of the phenomena at issue: namely, those cases where implicit biases are *dissonant* with explicit beliefs (cf D2). This is because, as Holroyd (2016) has argued, it is difficult to apply this view to cases in which implicit and explicit beliefs are aligned (is there then one belief, both conscious and unconscious?).

Notwithstanding the empirical evidence that Mandlebaum marshals in support of the unconscious belief view, some authors remain unconvinced of the claim that they are beliefs thus construed. For example, Madva (2016) has argued that implicit biases are probably not beliefs, since empirical evidence suggests a number of cases in which they fail to meet what he specifies as a necessary condition for belief, namely sensitivity to logical form. Similarly, Levy (2015) has suggested that even if biases have propositional structure, they appear to be insufficiently responsive to evidence to support the claim that they function as beliefs do. Rather, Levy proposes a *sui generis* mental state that better accommodates the characteristic features of implicit biases (see section b. below).

ii. in-between beliefs

Schwitzgebel (2010) has a rather different proposal for modelling implicit biases: as cases of 'in-between belief'. This assumes an understanding of beliefs as broad track dispositions. When we have dispositions - to assert, to behave - in ways that belie seemingly inconsistent beliefs, Schwitzgebel claims, we should say that we 'kind-of' or 'in-between believe'. For example, an individual who professes a commitment to racial equality, yet nonetheless under-evaluates the CVs of black and minority ethnicity applicants, has dispositions

[21] This survey is not comprehensive, in part because of the rate at which the literature is developing. For example, a number of philosophers are also exploring the idea, in unpublished work, that implicit biases are best understood as imaginings.

consistent both with believing that race is irrelevant to the suitability of an applicant for a job, and not believing this.

Importantly, this view does not take on any particular commitments about the sense in which implicit biases are *implicit;* and in fact no distinctive properties are attributed to the dispositions that manifest what we might identify as the implicit beliefs. Indeed, the analysis given in the case of implicit bias would not differ from a case in which an agent has conflicting explicit beliefs: in both cases we ascribe conflicting beliefs on the basis of dispositions that indicate the agent believes *p* and not-*p*. For those who have doubts about the feasibility of the distinction between implicit and explicit biases (cf. D1 vs. D4), this may be a virtue of the account. Second, note that this view appears motivated in particular by the concern to capture cases in which there is dissonance between the biases and agent's asserted beliefs (D2). In cases where an agent's biases accord with her beliefs and values, there will be no 'in-between' belief to speak of (see Holroyd 2016). The appeal of this account, then, may depend on the scope of the phenomena one is seeking to capture.

b. *sui generis* states

Some authors have suggested that the empirical findings about implicit bias cannot easily be made sense of within a framework that posits familiar, folk psychological mental states. Instead, new *sui generis* states need to be introduced to make good sense of the phenomena.

i. patchy endorsements

Levy (2015), for instance, has argued that implicit biases are best understood as mental states that he dubs 'patchy endorsements'. These mental states have some propositional structure, but are not responsive to reasons in the way that other of our attitudes, such as beliefs, typically are. The 'patchiness' of biases means that they do not stand in inferential relations with other attitudes, and so cannot integrate with the agent's other mental states. This has implications for their role in agency: patchy endorsements are not integrated into the agent, so they cannot be attributable to an agent (she cannot be blamed for them or their role in action).

Again, we can tease out the theoretical choices informing this view: first, it is largely driven by the empirical evidence - which suggests both that biases may sometimes have propositional structure, and that they are not involved in inferential processing to the same degree as other mental states (D4). Second, an underlying assumption of modelling implicit biases as 'patchy' appears to be that what distinguishes implicit attitudes is that they are not subject to the same kind of normative or rational control that other attitudes can be governed by (D1). But recall that we saw above ~~note also~~ the assumption that a lack of integration with other inferential states means that the biases are not integrated into the agent as a whole - and to the extent that they are, Levy suggests, their status as 'implicit' is in doubt. This indicates that the scope of the phenomena with which Levy is concerned, again, is primarily ~~restricted to~~ those instances~~cases~~ in which the implicit bias conflicts with the agent's other attitudes (D2). The account is less well placed to speak to cases in which the agent's biases are aligned with and reinforce her biased explicit attitudes (D4). If one maintains that these cases in which biases are aligned with the agent's beliefs and values constitute an important subset of the phenomena at issue (see Holroyd 2016, Zheng 2016), then this will be a problematic aspect of Levy's account.

ii. aliefs

Gendler has introduced the notion of *alief* to capture certain aspects of our cognition, including implicit biases. These mental states, aliefs, are associative, automatic and arational, and are activated by the agent's environment (2008a, p.642). Aliefs are states constituted by tripartite clusters of co-activated contents: this includes representational content, affective states, and the readying of motor responses. The representational content of aliefs need not be propositional, and aliefs can be held consciously or non-consciously. For example, the alief model would reconstruct implicit biases involved in CV studies as including the following contents: 'black applicant [representational]; negative affect/evaluation [affective response]; deflate judgement/avoidance response [readying of motor responses]'. Gendler uses discordant cases, where aliefs conflict with, and serve to undermine, the agent's endorsed values and attitudes, to elucidate the phenomenon. However, she emphasises that aliefs may also be in accordance with her beliefs and other explicit attitudes (2008b, p.554).

We can already see that unlike the two belief accounts, and the patchy-endorsement view, considered above, this account is positioned to capture a broader range of phenomena (D4). Aliefs may be involved not only in those cases where biases are dissonant with and undermine the agent's values; but also those in which the biases - aliefs - underpin and support the agent's explicit attitudes. Another feature of Gendler's treatment of biases is that in identifying them as aliefs, she locates them alongside a range of other phenomena that she describes as involving automatic and associative states, such as aversion to eating fudge shaped like faeces; fear responses to standing high up in locations one knows to be safe (2008a). In so positioning implicit biases qua aliefs, we see that the distinguishing feature of these phenomena is not that they are unconscious or inaccessible - one can perfectly well observe one's aliefs in relation to the high walk-way or faeces-shaped chocolate. Rather, a more likely candidate for distinguishing them from mental states involved in reflective or deliberative thought (D1) is rather that their activation is unintentional, or that their operation is not under our control, in ways other mental states may be. For example, Gendler suggests that the co-activation of the constituents of aliefs is automatic, such that the representational component will automatically activate the affective and behavioural components. In contrast, explicit beliefs are 'combinatoric', namely, apt for combination with any other belief or desire.

Some authors have expressed scepticism about the notion of alief (e.g. Currie & Ichino 2012, discussed in Holroyd 2016), and suggested that our picture of the mind can accommodate the constituents of alief (affect, representation, motor response) without supposing that they cluster and constitute a new sort of mental state. Indeed, there may be reasons to avoid a model of implicit bias that maintains they uniformly have this tripartite feature, since there may be various dimensions of heterogeneity in the mental states that comprise the phenomena (see ~~Madva & Brownstein 2016,~~ Holroyd & Sweetman 2016, for discussion of these dimensions of heterogeneity[22]). If one seeks a model that can capture all of these phenomena (D4), then modelling implicit bias as tripartite aliefs may be problematic.

[22] Madva & Brownstein 2016 also identify important dimensions of heterogeneity that an account of implicit biases should accommodate, but note that they see their view as compatible with seeing biases as 'alief-like'.

c. traits

The views we have considered so far all posit implicit biases are some sort of mental state - beliefs (variously characterised), or some *sui generis* mental state such as a patchy endorsements, or aliefs. This supposes that implicit biases can be tokened implicitly or explicitly: any such view thereby takes on a commitment to explicating what it is to be implicit rather than explicit (D1). This assumption has been put under pressure by Machery, who has suggested that we should think of attitudes as traits. As such, attitude are dispositions to cognize, respond affectively, and behave in certain ways. Since they are dispositional profiles, they are not the sort of thing that can be implicit or explicit. What have been referring to as implicit biases are the manifestation of certain dispositions to evaluate, or feel certain ways, in response to certain stimuli. The sorts of things we might be inclined to call 'explicit attitudes', Machery argues, are better understood as judgements about attitudes, rather than expressions of attitudes themselves. These judgements can be more or less accurate, which is what accounts for the cases in which there is dissonance between an agent's behaviour or affective response, and her pronouncements about what she believes or values (cf. D2).[23]

This view explicitly seeks to avoid any commitment to identifying the distinguishing property between implicit and explicit attitudes, then (cf. D1). Moreover, such a view fares better in accommodating the empirical evidence (D4), Machery argues: the complexity of our attitudes, construed as traits, explains why there is little correlation between indirect measures, and why behaviour may not be predicted by the results of any one indirect measure. Since the various psychological bases of our attitudes are involved in the production of behaviour, we would not expect any one such component to predict it.

However, note that the success of this view trades on an ambiguity in the literature to which we have not yet adverted: that between implicit bias and implicit attitudes. Some authors refer to the phenomena at issue as implicit bias; others to implicit attitudes (in particular, Levy 2014, 2017~~forthcoming~~). This elision occurs partly because psychologists refer to evaluative responses as attitudes, and partly because one of the points of contention in the philosophical literature is whether implicit biases can be attributable to the agent, and therefore be part of the evaluation of 'who the agent is' and 'what she stands for': issues that have typically been referred to as pertaining to the agent's attitudes (cf. Holroyd 2012, Brownstein 2015, Levy 2014, Zheng 2016). This matters for Machery's view, since he is suggesting that the agent's various dispositions constitute her attitudes, and these attitudes, understood as multi-track dispositional profiles, cannot be implicit or explicit. However, this view still supposes that there are various psychological bases of the attitude - some of which may be mental states, and some of which may therefore admit of being implicit or explicit. If biases at issue in the phenomena we have been discussing here are identified with some of *these* mental states (which are constituents of attitudes), then it will remain a live question as to what sense, if any, *they* are implicit, and what property (if any) distinguishes them from explicit mental states that also constitute the psychological basis of the disposition.

[23] Machery's trait view may have difficulties ascribing attitudes to agents whose dispositional profiles reflect ambivalence of attitude (since their dispositions reveal neither an all things considered positive, nor all things considered negative, attitude). Lee proposes an alternative conceptualisation that accommodates both the mean and distribution of likings/dislikings and other attitude relevant states, and as such, she argues, is better able to deal with such ambivalence. See Lee (ms.).

e. eliminativism

In this section, we introduce an alternative view that, as far as we know, is yet to be argued for. This view is eliminativism about implicit bias. It holds that there is no such psychological kind, and therefore no account that attempts to characterise implicit bias as a particular mental state or psychological kind will succeed. On this view, there is no unified phenomena, with any distinctive set of characteristics, that underpins the behavioural responses found on indirect measures such as those we introduced at the start. Rather, a cluster of different mental states and processes may produce these responses; and these mental states and processes may also be involved in the production of responses on other measures, such as self-report or direct measures.

This has much in common with Machery's proposal that the psychological basis of traits is diverse, which is to say that an agent's dispositions to respond to social groups will include cognitive associations, affective responses, propositional attitudes, evaluations, motor responses and so on. Yet, eliminativism resists the temptation to unite these responses into a psychological kind, 'attitude' (or anything else).[24] Instead, eliminativism maintains that the phenomena of implicit biases is best understood as involving various different mental states and processes, which may not share any property other than that of being recorded by a particular indirect measure (but this is not definitive of the implicit, since the states measured may also be recorded by self-report or other measures).

Presumably the appeal of such a view would be that its flexibility positions it well to capture the phenomena at issue (D4), including the heterogeneity of the phenomena, in terms of the states involved and the different kinds of behavioural outputs they might be implicated in (see Holroyd & Sweetman 2016; Madva & Brownstein 2016). It remains to be seen whether such a view is supported by empirical evidence, and the details will be important in ascertaining whether it is able to deliver adequate interventions (D3). But note that such a view need not commit to any particular property as characteristic of or necessary for a state being implicit (rather than explicit) (D1). Rather, this view could maintain that this is simply a convention used to refer to those states or processes involved in the production of responses on indirect measures. From the point of view of constructing an adequate picture of the mind, it is somewhat misleading to refer to these states as 'implicit biases', as if that identified a psychological kind. From the perspective of constructing an adequate theory of our psychological goings on, we could eliminate the notion of 'implicit bias' and still be able to well describe what states and processes are involved in the responses measured.

One might resist this move by insisting that the notion of implicit bias can pick out a psychological kind, but specify a more fine-grained notion than that which has so far been at issue: a subset of those biases that have been identified as implicit biases. Such a revisionary view could then be evaluated in terms of the desiderata it satisfies.

Does taking an eliminativist view about the psychological reality of implicit bias mean that one must also be eliminativist about the language and terminology of 'implicit bias' and that we would do better to wholesale avoid such misleading references? Recall that one of the desiderata outlined above was primarily pragmatic (D5). If one seeks to draw attention to the phenomena of discrimination that has long been overlooked or ignored and motivate

[24] Of course, to suggest eliminativism about implicit bias and deny that they are e.g. attitudes is not to suggest eliminativism about attitudes.

collective responsibility in addressing it, then it matters not if there is no unified psychological kind involved in those discriminatory patterns.

This seems to us a legitimate usage of the notion of 'implicit bias', despite its commitment to the absence of any such mental item as 'implicit bias' - so long as proponents of this view are clear that they are motivated by aims other than to capture the psychological reality of our minds. However, if one has doubts about the efficacy of the notion even for these political purposes, then a wholesale eliminativism may have appeal.

7. Concluding remarks

We have surveyed some of the competing views about the domain of the implicit, the nature of bias, and the psychological reality of implicit biases. We sought to identify the often unarticulated aims or tacit desiderata that each position appears to be motivated by, and teased out the commitments of each view. We do not here intend to take a decisive stance on the merits of these aims or commitments, but merely to identify that there *are* such commitments, since these are theoretical choices about which authors have not always been clear. Our hope is that bringing to light some of these competing aims enables a clearer and more robust defence of the notion of implicit bias, and a clearer sense of what we want to use that notion for.

References

Amodio, D. M. and Devine, P. G. (2006). "Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior." *Journal of Personality and Social Psychology* 91: 652–61.

Antony, L (1993) "Quine As Feminist: The Radical Import of Naturalized Epistemology," *A Mind of One's Own: Feminist Essays on Reason and Objectivity*, Louise Antony and Charlotte Witt (eds.), Boulder: Westview Press, 185–226.

Antony, L. (2016) "Bias: Friend or Foe?", in Brownstein, M. & Saul, J. (eds.) *Implicit Bias and Philosophy, volume 1*, Oxford: Oxford University Press, 157-190.

Bacchus, J. R., Baldwin, M. W., & Packer, D. J. (2004). "Increasing implicit self-esteem through classical conditioning". *Psychological Science*, 15, 498–502.

Banaji, M. R. (2001). "Implicit attitudes can be measured". In H. L. Roediger, J. S. Nairne, I. Neath, & A. Surprenant (Eds.), *The nature of remembering: Essays in remembering* Robert G. Crowder (pp. 117–150). Washington, DC: American Psychological Association.

Banaji, M. & C. Hardin, 1996, "Automatic stereotyping", *Psychological Science*, 7(3): 136–141.

Bertrand, M. & S. Mullainathan, 2004, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market", NBER Working Papers from National Bureau of Economic Research, Inc., No. 9873

Blum, L. (2004) "Stereotypes and Stereotyping: A Moral Analysis". *Philosophical Papers* 3 (2004): 251–89.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). "Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited?" *Journal of Personality and Social Psychology,* 79, 631–643.

Brownstein, M. & Saul, J. (2016) "Introduction", in Brownstein, M & Saul, J. (eds.) *Implicit Bias and Philosophy volume 1* Oxford, Oxford University Press, 1-19.

Brownstein, M. (2015) "Attributionism and Moral Responsibility for Implicit Bias", *Review of Philosophy and Psychology* doi:10.1007/s13164-015-0287-7 pp.1-22

Brownstein, Michael, "Implicit Bias", The Stanford Encyclopedia of Philosophy (Spring 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2017/entries/implicit-bias/>.

Brunstein, J. C., & Schmitt, C. H. (2004). "Assessing individual differences in achievement motivation with the Implicit Association Test". *Journal of Research in Personality*, 38, 536–555.

Carruthers, P. (2010). "Introspection: divided and partly eliminated". *Philosophy and Phenomenological Research*, 80, 76–111.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.

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.

Currie, G. & Ichino, A. (2012) "Aliefs don't exist, but some of their relatives do", *Analysis*, 72: 788–798.

Devine, P.G., Plant, E.A., Amodio, D.M., Harmon-Jones, E. & Vance, S.L., (2002) "The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice," *Journal of Personality and Social Psychology* 82 (2002): 835–48.

Dotson, K. (2011) "Tracking Epistemic Violence, Tracking Practices of Silencing", *Hypatia* vol.26(2) pp.236-257.

Dovidio, J. F., & Gaertner, S. L. (2000). "Aversive racism and selection decisions: 1989 and 1999". *Psychological Science*, 11: 319–323.

Dovidio, J. F., Kawakami, K & Gaertner, S. E, (2002) "Implicit and Explicit Prejudice and Interracial Interaction" *Journal of Personality and Social Psychology*, Vol. 82, No. 1, 62–68

Dovidio, J. F., Gaertner, S. E., Kawakami, K., & Hodson, G. (2002). "Why can't we just get along? Interpersonal biases and interracial distrust", *Cultural Diversity and Ethnic Minority Psychology*, *8*(2), 88-102.

Dovidio et al, Kawakami, Johnson & Johnson, (1997) "On the nature of prejudice: Automatic and controlled processes" *Journal of Experimental Social* Psychology, 33, 510-540;

Duguid, M. M. & Thomas-Hunt, M.C. (2015) "Condoning stereotyping? How awareness of stereotyping prevalence impacts expression of stereotypes". *The journal of Applied Psychology,* 100(2):343-59.

Faucher 2016, "Revisionism and Moral Responsibility for Implicit Attitudes" in Brownstein, M., & Saul, J. (eds.) *Implicit Bias and Philosophy volume 2* pp.115-145

Fazio, R. H., & Olson, M. A. (2003). "Implicit measures in social cognition research: Their meaning and uses". *Annual Review of Psychology*, 54, 297–327.

Fazio, R. & T. Towles-Schwen, 1999, "The MODE model of attitude-behavior processes", in Chaiken & Trope (eds.) *Dual-process theories in social psychology*, New York: Guilford Press, pp.97–116.

Follenfant, A. & F. Ric, 2010, "Behavioral Rebound following stereotype suppression", *European Journal of Social Psychology*, 40: 774–782.

Frankish, K. (2016) "Playing double: Implicit bias, dual levels, and self-control", in Brownstein, M., & Saul, J. (eds.) *Implicit Bias and Philosophy, volume 1*, Oxford: Oxford University Press, 23-46.

Garcia, J. L. A. (2004) 'Three Sites for Racism: Social Structures, Valuings and Vice'. *Racism in Mind.* Eds. M. P. Levine and T. Pataki. Ithaca, NY: Cornell UP, pp.36–55

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

Gendler, T. (2008a) "Alief and Belief" *Journal of Philosophy* 105 (10): 634-663

Gendler, T. (2008b). "Alief in Action (and Reaction)". *Mind and Language* 23 (5):552--585.

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

Glasgow, J. (2016) "Alienation and Responsibility" in Brownstein, M., & Saul, J. (eds.) *Implicit Bias and Philosophy volume 2* pp.37-61.

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

Greenwald, A, G, Nosek, B, A, Banaji, M, R. (2003) "Understanding and using the Implicit Association Test: I. An improved scoring algorithm". *Journal of Personality and Social Psychology*, 85(2):197–216.

Gregg, A. P. & Klymowsky, K. (2013) "The Implicit Association Test in Market Research: Potentials and Pitfalls" *Psychology & Marketing* , 30 (7): 588–601

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2013). "Awareness of implicit attitudes". *Journal of Experimental Psychology*: 143(3): 1369-92

Hall, L., Johansson, P., & Strandberg, T. (2012). "Lifting the veil of morality: choice blindness and attitude reversals on a self-transforming survey". *PLoS One*, 7(9):e45457. doi: 10.1371/journal.pone.0045457

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

Haslanger, S. (forthcoming 2017) "Racism, Ideology and Social Movements" *Res Philosophica*

Holroyd, J. (2012) "Responsibility for Implicit Bias" *Journal of Social Philosophy, Special Issue: Gender, Implicit Bias and Philosophical Methodology* Crouch, M. & Schwartzman, L. (eds.) 43(3), pp. 274–306

Holroyd, J. (2014) "Implicit Bias, Awareness and Imperfect Cognition" in *Consciousness and Cognition, Special Issue: Costs and Benefits of Imperfect Cognitions* Bortolotti, L. & Sullivan-Bissett, E. (eds.) pp. 511-523

Holroyd, J. (2016) "What do we Want from a Model of Implicit Cognition?" *Proceedings of the Aristotelian Society* 116(2) pp.153-179

Holroyd, J & Kelly, D. (2016) "Implicit Bias, Character and Control" in Webber, J. and Masala, A. (eds.) *From Personality to Virtue: Essays on the Philosophy of Character* Oxford University Press, pp.106-134

Holroyd, J. & Puddifoot, K. (forthcoming) "Implicit Bias and Prejudice" *Routledge Handbook of Social Epistemology*, edited by Miranda Fricker, Peter J. Graham, David Henderson, Nikolaj Pedersen, and Jeremy Wyatt.

Holroyd, J. & Sweetman, J. (2016) "The Heterogeneity of Implicit Biases", in Brownstein & Saul (eds.) *Implicit Bias and Philosophy volume 1* Oxford: Oxford University Press, 80-103

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). "Failure to detect mismatches between intention and outcome in a simple decision task". *Science*, 310, 116–119. 730

Jost, J. T., Pelham, B. W., & Carvallo, M. R. (2002). "Non-conscious forms of system justification: Implicit and behavioral preferences for higher status groups". *Journal of Experimental Social Psychology*, 38, 586–602.

Jost, J.T., Rudman, L., Blair, I.V., Carney, D.R., Dasgupta, N., Glaser, J., & Hardin, C. (2009). "The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore.: *Research in Organizational Behavior*, 29, 39-69.

Kelly, D. (2013). "Implicit Bias and Social Cognition", *The Encyclopedia of Philosophy and Social Science*, Ed. B. Kaldis. Thousand Oaks, CA: SAGE Publications. Vol 9: 460 - 462. doi: http://dx.doi.org/10.4135/9781452276052.n172

Kelly, D. and Roedder, E. (2008). "Racial Cognition and The Ethics of Implicit Bias", *Philosophy Compass*, 3/3, April 2008, pages 522 - 540. doi:10.1111/j.1747-9991.2008.00138.x

Lee. C. (ms.) "A Dispositional Account of Aversive Racism".

Levy, N. (2015) "Neither fish nor fowl: Implicit attitudes as patchy endorsements", *Noûs*. doi:10.1111/nous.12074

Levy, N. (2017) "Implicit Bias and Moral Responsibility: Probing the Data" *Philosophy and Phenomenological Research*, 94 (1): 3-26 doi: 10.1111/phpr.12352

Macy, J.T., Chassin, L, Clark, C.P. (2013) "The Association Between Implicit and Explicit Attitudes Towards Smoking and Support for Tobacco Control Measures", *Nicotine and Tobacco Research* 15 (1): 291-296.

Machery, E., Faucher, L. & D. Kelly, 2010, "On the alleged inadequacy of psychological explanations of racism", *The Monist*, 93(2): 228–255.

Machery, E. (2016) "De-Freuding Implicit Attitudes", in Brownstein & Saul (eds.) *Implicit Bias and Philosophy volume 1* Oxford; Oxford University Press 104-129.

Madva, A. (2015) "Why Implicit Attitudes are (probably) not beliefs" Synthese, pp.1-26 DOI 10.1007/s11229-015-0874-2

Madva & Brownstein (2016) “Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind”. *Noûs* 50 (4).

Mandlebaum, E. (2015) “Attitude Inference and Association: On the Propositional Structure of Implicit Bias” *Nous* 10.1111/nous.12089

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., & Handelsman, J. (2012). “Science faculty’s subtle gender biases favor male students”. *Proceedings of the National Academy of Sciences*, 109, 16474-16479.

Nosek, B. A., A. G. Greenwald, and M. R. Banaji. (2007) “The Implicit Association Test at Age 7: A Methodological and Conceptual Review’ *Automatic Processes in Social Thinking and Behavior*. Ed. J. A. Bargh. Philadelphia, PA: Psychology Press, 265-292.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). “Harvesting implicit group attitudes and beliefs from a demonstration website”. *Group Dynamics*, 6(1), 101-115.

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

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.

Phelps, E. A., O’Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., et al. (2000). “Performance on indirect measures of race evaluation predicts amygdala activation”. *Journal of Cognitive Neuroscience*, 12, 729–738.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. K. (1999). “Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test”. *Social Cognition*, 17, 437–465.

Saul, J. (2013a) “Implicit Bias, Stereotype Threat, and Women in Philosophy” in Jenkins, F. & Hutchinson, K. *Women in Philosophy: What needs to change?* Oxford University Press.

Saul, J. (2013b) “Implicit Bias and Scepticism” *Disputatio* 5: 37, 243-263.

Scaife, R. Holroyd, J., Stafford, T., Bunge, A., (ms.) “The Effects of Moral Interactions on Implicit Racial Bias” https://osf.io/eubjp/

Schwitzgebel, E. (2010) “Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief”, *Pacific Philosophical Quarterly*, 91: 531–553.

Spalding, L. R., & Hardin, C. D. (1999). "Unconscious unease and self-handicapping: Behavioral consequences of individual differences in implicit and explicit self-esteem." *Psychological Science*, 10, 535–539.

Stacy, A.W., Newcomb, M.D. & Ames, S.L. (2000) Journal of Behavioural Medicine 23 (5): 475–499

Steinpreis, R.E., Anders, K.A. & Ritzke, D. (1999) "The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study" *Sex Roles*, 41: 509. doi:10.1023/A:1018839203698

Teachman, B. A., Gregg, A. P., & Woody, S. R. (2001). "Implicit associations for fear-relevant stimuli among individuals with snake and spider fears". *Journal of Abnormal Psychology*, 110, 226–235.

Wilson, T. (2002). *Strangers to Ourselves*. Harvard University Press.

Zheng (2016), "Attributability, Accountability and Implicit Attitudes", in Brownstein, M. & Saul, J. (eds.) *Implicit Bias and Philosophy* vol.2 pp.62-89

Internet sources:

The Brains Blog: What can we learn from the Implicit Association Test? A Brains Blog Roundtable http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx accessed 11/02/2017

Project Implicit FAQ: https://implicit.harvard.edu/implicit/faqs.html#faq0 accessed 11/02/2017

Wikipedia entry on 'Implicit Attitude': https://en.wikipedia.org/wiki/Implicit_attitude accessed 11/02/2017

Hillary Clinton talks race: 'We all have implicit biases' Dan Merica, CNN April 21 2016, **http://edition.cnn.com/2016/04/20/politics/hillary-clinton-race-implicit-biases/index.html** accessed 11/02/2017

Clinton: 'I Think Implicit Bias Is a Problem for Everyone', Melanie Arter, CNS News, Sept 27 2016, http://www.cnsnews.com/news/article/melanie-hunter/clinton-i-think-implicit-bias-problem-everyone accessed 11/02/2016