
Article:
Holroyd, J.D. (2016) What Do We Want from a Model of Implicit Cognition? Proceedings of the Aristotelian Society, 116 (2). pp. 153-179. ISSN 0066-7374

https://doi.org/10.1093/arisoc/aow005

OXFORD UNIVERSITY PRESS | Proceedings of the Aristotelian Society

What Do We Want from a Model of Implicit Cognition?

Journal:	*Proceedings of the Aristotelian Society*
Manuscript ID	Draft
Manuscript Type:	Invited Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Holroyd, Jules; University of Sheffield, Department of Philosophy
Keywords:	implicit bias, implicit cognition, belief, alief, automaticity, racism

What do we want from a model of implicit cognition?

Jules Holroyd

Abstract: In this paper, I set out some desiderata for a model of implicit cognition. I present test cases and suggest that, when considered in light of them, some recent models of implicit cognition fail to satisfy these desiderata. The test cases also bring to light an important class of cases that have been almost completely ignored in philosophical discussions of implicit cognition and implicit bias. These cases have important work to do in helping us understand both the role of implicit cognition in action, and our attempts to combat implicit biases.

In this paper, I set out some desiderata for a model of implicit cognition. I present test cases, and suggest that, when considered in light of them, some recent models of implicit cognition and in particular, of implicit bias - from Levy, Schwitzgebel, Mandelbaum, Gendler and Machery - fail to meet these desiderata. I propose a 'minimal model' that may fare better. I also bring to light an important class of cases that have been almost completely ignored in philosophical discussion of implicit cognition and implicit bias. These cases have important work to do in helping us understand both the role of implicit cognition in action, our responses to implicit bias and attempts to combat it.

Models of implicit cognition have emerged from different sources with different priorities and conceptual frameworks. Psychologists have developed such models as a way of understanding what processes or mechanisms best explain the experimental findings about implicit associations and their impact on behaviour. Meanwhile, recent philosophical accounts of implicit cognition have focused on implicit bias in particular. These accounts have been formulated with various aims in mind, including: incorporating implicit bias into a complete picture of our psychological states and their role in action; articulating the role of these implicit cognitions in a normative conception of agency or responsibility; identifying the epistemic implications of this aspect of our agency. To this end, it is no surprise that philosophical conceptions of implicit bias vary significantly. But how might we go about evaluating whether a model of implicit bias is successful? And how do the accounts developed by philosophers, to date, fare? My task in this paper is to start to address these questions.

In section I, I set out some desiderata for a model of implicit cognition. The desiderata I set out are not intended to be exhaustive, but provide a starting framework for evaluating models of this aspect of our cognition. In section II, I introduce some test cases which can be used to evaluate whether the accounts of implicit bias satisfy our desiderata. These cases also bring to the forefront some cases of implicitly biased action that have been almost wholly absent from the philosophical discussion of implicit bias. In section III, I then evaluate five views which provide models for understanding what some of our implicit cognitions - implicit biases - are. Levy claims that implicit biases are patchy endorsements; Schwitzgebel claims they are in-between beliefs; Mandelbaum suggests they are unconscious beliefs; Gendler models them as *aliefs;* Machery argues that implicit attitudes properly understood are not implicit at all. I evaluate how well each view is able to satisfy the desiderata from section I, and find them wanting. They cannot accommodate our test cases from section II, nor capture the distinctive explanatory and moral significance of implicit cognition that we observe in these cases. I close with a proposal for an account that might more satisfactorily meet these desiderata.

I

Desiderata for a model of implicit cognition. The sort of phenomena that models of implicit bias try to capture are by now familiar: the processes or states that have a distorting influence on behaviour and judgement, and are detected in experimental conditions with implicit measures, such as the

implicit association test. The most troubling of these concern distortions to do with social identity: gender, race, age or sexuality biases, for example. Such implicit cognitions are typically understood as fast and automatic processes, which may operate without the agent being aware of their activation or influence on behaviour. For example, implicit racial biases have been found to have a role in influencing the evaluation of CVs - participants rated identical CVs less positively when the name at the top indicated that it was the CV of an individual racialised as black (Dovidio & Gaertner 2000). Implicit biases may manifest in 'micro-behaviours' - unintentional non-verbal behaviours that indicate tension or discomfort. White people are found to display these in interracial interactions with black interlocutors, affecting the quality of the interpersonal interactions (Dovidio, Kawakami & Gaertner 2002; Dovidio, Gaertner, Kawakami & Hodson 2002). And, disturbingly, implicit associations between black people and weapons are found to affect performance on shooter bias tasks, with individuals more likely to make the error of shooting an unarmed individual when that individual is black (Glaser & Knowles 2007).

Since these judgements and behaviours are poorly described by more familiar psychological states - such as the agent's explicit beliefs, intentions and values - we need an account of what it is that produces such discriminatory behaviour. Accounts of implicit bias aim to capture the parts of our psychology implicated in these behaviours. Such an account aims to articulate what this part of our psychology consists in; what it does; and how it is integrated with other cognitions. The inquiry is complicated by the fact that there is no consensus regarding what it is that makes a part of our mental workings 'implicit' rather than explicit. It is also complicated by the fact that a broad range of phenomena may fall under the rubric of implicit cognition, and it may be that there are different understandings of 'implicit' involved in these phenomena. Our present focus is on models of implicit bias. In evaluating accounts of implicit bias, we should at this stage leave open both the question of what is distinctively 'implicit' about such cognitions, and whether an adequate account of implicit bias is continuous with other aspects of our implicit cognitions. With this in mind, we begin by asking what we want such an account to do.

(i) Empirical validity. Our theory should be sensitive to the findings of empirical psychology. A wealth of empirical data can be drawn on to construct our theory of implicit cognition. For example one dimension of the debate has focused on whether implicit attitudes have propositional or associative structure (Mandelbaum 2015, Levy 2015). The theory best placed will be (inter alia) the one best supported by the empirical evidence about whether implicit biases seem to operate in a way that only states with propositional rather than associative structure do - for example, whether they appear to be implicated in inferential patterns of reasoning or not. More generally, we should expect that the preferred model of implicit cognition finds support in the empirical literature. There are two dimensions of empirical validity in particular that are worth highlighting, as follows.

(ii) Bias in context. We know that implicit biases are, in part, caused by the associations we find in our cultural and social environment - such that race biases are contingent upon a specific cultural context that encodes racist stereotypes (Devine 1989). We also know that implicit biases can be reigned in by institutional structures, and biases unlearned (or new, non-stigmatising biases learned) in the appropriate social contexts.[1] Accordingly, our model of implicit cognition should cohere with these findings and should elucidate our understanding of how implicit cognitions are shaped by these social and institutional structures.

(iii) Framework for understanding how to change implicit cognitions. Since our focus is on implicit biases that concern the stigmatisation of social identities, one important aspect of a theory of implicit cognition is that it should help us to understand how we might change or combat implicit

1 See Jost et al 2009 for an overview of institutional contexts in which biases might operate; and Lai et al 2014, for changes that may prevent biases from operating. See also 'Beyond Bias and Barriers' by the NAC for specific proposals for institutional change (in the context of STEM research).

biases. The theory should deliver hypotheses that generate testable predictions about what might change implicit biases, or how their expression might be mitigated. This is not to say that the model must deliver prescriptions for changing implicit biases; but rather that a useful model will generate hypotheses that we can test. It is consistent with this that when tested, no interventions are found to be successful. This at least helps us to see where we should not allocate our efforts. The point is that if a theory can generate predictions, the confirmation of those predictions provides additional empirical validity. Moreover, it is desirable that a theory of implicit bias aid us with the transformative work needed to remedy problematic implicit biases.

However, the empirical findings are often consistent with more than one interpretation of the data, and under-determine which model we might endorse. So we need further desiderata to guide us.

(iv) Coherence with our understanding of cognition in general. We should want our model of implicit bias to integrate with our best understanding of cognition in general. This is not to say that our theories of implicit cognition should be constrained to utilise only those notions or posits of extant theories of mind. It may be (we shall indeed see examples of this shortly) that new sorts of state, or revisionary understandings of existing states, have to be posited in order to make good sense of the implicit workings of our psychologies. But we should be in a position to explain how those workings can be integrated into a fuller picture of our minds. One hope is that we get an account of implicit bias that is able to explain how implicit biases relate to, and interact with, explicit cognitive processes or explicit mental states. This requires also that our model elucidates what it is for this aspect of cognition to be *implicit* and how it is distinguished from explicit aspects of cognition. This is no small task, of course - and as we shall see, there are problems with extant understandings of this.

(v) Framework for understanding the role of implicit cognition in moral agency. Since many philosophers engaged in this debate are not only interested in what is in our minds, and how that influences behaviour, but also for what this means for beings who are moral agents, we should seek a theory of implicit cognition that coheres with a framework for understanding how this aspect of our cognitive workings is implicated in our moral agency. This is not to say that the theory should be constrained by folk psychological notions of what moral agency consists in. Moreover, we should remain open to our judgements about moral agency and moral responsibility needing considerable refinement or revision in order to be suitably responsive to empirically informed understandings of our agency. It may be that our best account - along with assumptions about the nature of moral agency, responsibility and evaluation - delivers the judgement that implicit biases are not the sort of thing for which we can be morally evaluated. But our understanding of moral agency and moral responsibility provides some useful points of reference for evaluating models of implicit bias.[2] Suppose our judgements about agents' responsibility clearly distinguishes morally X and Y. But our model of implicit bias fails to distinguish between the moral status of X and Y. This gives us some reason both to consider revising our judgements about X and Y, and to check whether our model of implicit bias rests on defensible assumptions, proper interpretation of empirical findings, and so on.[3]

There may be other desirable features of a model of implicit cognition. I do not claim these to be exhaustive. However, I take these to be useful desiderata when evaluating such models. I next

2 In articulating the desiderata for an analysis of discrimination, Lippert-Rasmussen suggests that an analysis should be 'morally enlightening', in helping us to tease out moral conclusions (if not entailing them) (2014, 48). This is what I have in mind here.

3 There are some respects in which the project of articulating desiderata resonates with Cudd's (2005) articulation of desiderata with which to evaluate a theory of oppression. However, note that my desiderata different in some important respects: unlike Cudd's criteria of axiological pragmatism (p.32), I do not require that the theory tell us how to change problematic biases (but merely generate predictions for this). Likewise, the desiderata regarding a moral framework does not make any claim about the primacy of the value of moral agency (as Cudd does, p.36).

delineate a set of test cases. These cases help us to see the importance of these desiderata; second, they bring to light a kind of case that has been neglected in the philosophical literature; and thirdly, we can use these cases to test whether the models of implicit bias meet these desiderata, in section III.

II

Test cases. A notable feature of the philosophical literature on implicit bias is that it has focused overwhelmingly on what I call 'conflict' cases of implicit biases. By this I mean cases in which an agent's explicit values, commitments or endorsements are in conflict with a posited implicit bias which influences her actions in ways contrary to those explicit values, commitments and endorsements.

For example Saul proceeds with her attention on implicit biases in 'those who explicitly and sincerely avow egalitarian views' (2013, p.41). Glasgow's discussion of responsibility and alienation takes as its main focus cases of implicit biases which are in conflict with the agent's endorsed values (2016) - hence the alienation. Washington and Kelly's discussion focuses on cases in which there is 'dissociation' between implicit and explicit attitudes (2016). And Levy's analysis of implicit attitudes centres on cases in which implicit attitudes come apart from the agent's endorsed explicit attitudes, and influence behaviour in unwanted ways (2014, forthcoming). Conflict cases have been the overwhelming focus of the philosophical literature.

There are good reasons for focusing on these cases. Firstly, they are startling to many of us, since we have fair-minded explicit values, and commitments to non-discrimination. So it is discomforting to find that we have implicit biases that might implicate us in discrimination. Secondly, pragmatically speaking, these cases are an important class to focus on - for those with explicitly anti-discriminatory values and commitments, such as ourselves, are likely to be well motivated to endorse and implement efforts to overcome the influence of implicit biases. And finally, conflict cases raise a host of interesting and important philosophical considerations: what is the moral status of such implicit biases since they conflict with professed evaluative endorsements? Is the agent responsible for them? And so on.

However, this focus misses out an important class of cases. The cases I have in mind are those in which agent's implicit biases are aligned with her explicit attitudes. Whilst these cases have not received much philosophical attention, they are by no means mere conceptual possibilities, but empirical realities. Indeed, we know that agents with greater explicit prejudice record stronger implicit biases on implicit measures: Devine et al (2002) found that implicit race biases were stronger in individuals who harboured more explicit racial prejudice (see also Glaser & Knowles 2007). Accommodating non-conflict cases is important if a theory is to meet our desiderata. Consider the following three cases of individuals and their actions:

> *Racist*: Individual A has explicitly prejudiced beliefs and negative evaluations about black people. Further (and in accordance with what would be predicted on the basis of Devine et al's (2002) studies) this individual also harbours implicit negative associations. When on a hiring committee, A's evaluations are guided by his explicit prejudices (he does not want a black colleague, he believes that race is relevant to the suitability of a candidate). His behaviour is also influenced by his implicit biases, and he unintentionally displays certain 'micro-behaviours' in interracial interactions: he manifests discomfort, reacts with more irritability, sits marginally further away from the black interviewees (Dovidio. Kawakami, Gaertner & Hodson 2002, McConnell & Leibold 2001). He is not aware of these aspects of his behaviour, though he would not disavow them if he were. Unsurprisingly, A's makes an intentionally discriminatory decision: between the two equally well qualified candidates, the white rather than black candidate is hired.
>
> *Protocol adhering racist*: Individual B has explicitly prejudiced beliefs and negative evaluations about black people. Further (and in accordance with what would be predicted on

the basis of Devine's studies) this individual also harbours negative implicit associations. B also is strongly motivated in his professional role, and is very concerned to adhere to protocol, including fair hiring procedures. And indeed, if the best person to do the job is black, he thinks, they should be hired (still, his racism manifests in his preferences against going for after work drinks or invite anyone 'like that' round for dinner). B forms the intention not to discriminate when given responsibilities on a hiring committee. B's evaluations of the applicants are guided by his explicit intention not to discriminate. However, his behaviour is also influenced by his implicit biases: like A, he displays unintentional micro-behaviours: he manifests more discomfort, reacts with more irritability, and sits marginally further away from the black interviewees. He hasn't reflected much and is not aware of these aspects of his behaviour. Whilst these unintentional aspects of behaviour aligned with his explicit racist attitudes, they do not accord with his intention to adhere to protocol. These micro-behaviours affect the quality of the interaction, and when deciding between the two best qualified candidates, on the basis of the interview performance inflected by B's implicit biases, B judges there to be a clear margin between them - and the equally well qualified black candidate is not hired.

Biased egalitarian: This by now familiar individual, C, has explicitly anti-racist beliefs and egalitarian commitments. However, this individual also harbours implicit negative associations. When on a hiring committee, he intends not to discriminate (in accordance with his explicit anti-racist commitments). However, his implicit race associations mean that, like A and B, his behaviour is influenced by his implicit biases, and he manifests various micro-behaviours: he displays more discomfort, reacts with more irritability, and sits marginally further away from the black interviewees. These micro-behaviours affect the quality of the interaction, and when deciding between the two best qualified candidates, the black candidate is not hired.

Considering these three cases alongside each other helps us to see the role of our desiderata. First, an empirically valid account of implicit bias should be able to accommodate both conflict (C) and non-conflict cases (A and B), if they are to meet the desiderata of empirical validity: in each of these cases, implicit bias is implicated in discriminatory behaviour. Secondly, these cases bring out the differential first pass moral evaluations we might make of each individual. Each individual discriminates, but the causal and moral story behind this differs: A has morally repugnant attitudes and values, and intentionally discriminates. B too has repugnant explicit attitudes, but the fact that his discrimination is not intentional makes an important difference in the judgement of his behaviour.[4] C marks the paradigm 'conflict case' with which we are familiar from the philosophical literature: it is the conflict between the commitments and the discriminatory actions due to implicit bias that has garnered so much philosophical intrigue and attention. C too unintentionally discriminates. Some might not see a great deal of moral difference between agents in B and C, insofar as both unintentionally discriminate. But there are reasons to suppose that the agent in B is guilty of moral failures that the agent in C is not. Aside from the repugnant moral views B holds, he is also guilty of further moral and epistemic failures. Like A, even in the absence of knowledge about implicit bias, it is reasonable to expect that racist explicit beliefs will inflect automatic aspects of behaviour. B is guilty of culpable ignorance, or negligence in supposing that his behaviours would not be inflected by these morally problematic attitudes. On the other hand, an important difference between A and B is that whereas A would endorse these micro-behaviours were he aware of them, it is less clear that B would do so - despite their convergence with his explicit racism, they

4 I am assuming that implicit biases have a role in B's action. Why think this, rather than that a different mechanism is at work than in C, where there is no explicit prejudice? Since we find implicit bias prior to, and in the absence of, explicit biases, and since we know that various automatic cognitions run alongside our deliberative and explicit cognitions, it would not be parsimonious to posit a different mechanism in the case of B, and to suppose that implicit bias does not have any role in the discriminatory outcome here. I thank Zoe Drayson for pressing me on this point.

thwart other explicit intentions he holds. So, we can see the importance of a theory of implicit cognition delivering a framework that can make sense of these differential judgements.

These test cases also drive home the importance of a theory that can provide hypotheses about differential interventions regarding implicit biases. Restricting our attention to interventions that would tackle the racist discrimination manifested in each context, we can see that each case may require different interventions. For individual A, the racist, we have to address explicit as well as implicit racism; for the protocol adhering racist, B, strategies are needed that help to stop the bias taking effect (given the likelihood that the bias might persist due to its accordance with explicit prejudices); for the biased egalitarian, strategies would be welcome that insulate from the bias, and presumably the individual would also welcome strategies that change the implicit association itself.

III

Extant theories. In this section, we use these test cases to evaluate whether extant models of implicit bias can satisfactorily meet the five desiderata.

(i) Patchy-endorsements. Let us first consider Levy's account of implicit attitudes. Locating his view in the context of the dispute over whether implicit attitudes are propositional or associative in structure, Levy argues that they are best understood as 'patchy-endorsements'. That is, they are mental states which may have some propositional structure, but are not responsive to reasons in the way that explicit beliefs typically are.[5] What follows from this, crucially, is that implicit attitudes do not stand in inferential relations with other attitudes that are attributable to the agent. They do not stand in these relations because of their 'patchy' structure, which means they are not inferentially integrated with a wide range of other attitudes that are attributable to the agent. There is evidence that at least sometimes, implicit attitudes are acquired notwithstanding their conflict with the agent's endorsed attitudes, and are not extinguished when this is recognised (Levy, forthcoming, p.19). Accordingly, there is reason to believe that implicit attitudes are not integrated in the right way with agent's other attitudes. This is important, Levy claims, because attitudes that are attributable to the agent make up her deliberative standpoint, and those attitudes central to this standpoint reflect her evaluative position. Accordingly, for Levy, agents should not be held accountable for action that is influenced by implicit attitudes insofar as it is true that 'had the agent's explicit attitudes controlled behaviour, the actions would have had a different moral character' (forthcoming, p.25).[6]

There are two important claims from Levy to keep in mind: that because implicit attitudes are patchy, they will not be well-integrated with the agent's endorsed attitudes; and secondly, that insofar as they are not integrated in this way, the moral character of action influenced by implicit attitudes will diverge from the moral character of actions had her explicit attitudes guided them. Recalling our desiderata, we can say that at first glance, Levy gives us a theory supported by empirical evidence about the acquisition and extinction of implicit attitudes, and the extent to which they are responsive to reasons. This view locates implicit attitudes in terms of extant posits in philosophy of mind, whilst holding such attitudes importantly distinct from them. This theory also provides a framework for making sense of the moral evaluations we might make of implicit attitudes when they influence action: a framework which exculpates agents when the counterfactual is true. We also get some understanding about how to change implicit attitudes: we cannot suppose that propositional reasoning will be successful, since these attitudes are patchy in their responsiveness to reasons.

What does Levy's view say about our set of test cases above? The crucial two cases for our

5 Levy notes that implicit attitudes might meet Fischer & Ravizza's (2000) conditions for reasons-responsiveness - namely, that a mechanism is moderately responsive to reasons (responsive in a somewhat patterned way across some possible worlds). He takes this as a reason to reject that standard as sufficient. I am not convinced by that, but grant it insofar as it motivates us to look more closely at the issue of attributability.

6 Earlier, Levy's statement of the counter-factual differs slightly: 'had the agent's explicit attitudes alone controlled their behaviour, the action would have lacked its actual moral character' (7). I think this statement of the counterfactual generates slightly different judgements - but it is the version stated in the text that Levy relies upon in his attention to non-conflict cases, so I follow him in this.

purposes are those in which the agents implicit and explicit attitudes are aligned (individual A and B). Levy himself does not say much about such cases, since his focus is mainly on conflict cases, where the implicit and explicit attitudes diverge. But he does remark that, in cases in which implicit and explicit attitudes are aligned (where the implicit attitude is 'annexed' to the agent and 'well integrated' with her endorsed attitudes) the counterfactual claim we saw above will be false. That is to say it will be false that 'had the agent's explicit attitudes controlled behaviour the action would have had a different moral character' (p.25).

We can already start to see that something is going wrong here. Consider, from our set of test cases above, the case of the protocol-adhering racist (individual B). Here we have an individual who (ex hypothesi) has aligned implicit and explicit attitudes: his implicit and explicit attitudes are both racist. His implicit attitudes influence his behaviour. On Levy's view, since the implicit attitudes are aligned with the agent's explicit attitudes, it is false that had B's explicit attitudes controlled behaviour, the action would have had a different moral character. But this is wrong: had B's explicit attitudes controlled the behaviour, it *would* have had a different moral character. As described, the action is one of unintentional discrimination due the implicit biases influencing behaviour. Had his explicit racist attitudes controlled behaviour, it would have been an intentionally discriminatory action. And had his explicit concern for protocol controlled his behaviour, it would have been a non-discriminatory action. Both of these are importantly different in moral character from an action that is unintentionally discriminatory. In the case of individual B, the relevant counterfactual claim at issue is true: the action *would* have a different moral character if guided by explicit attitudes. But we might nonetheless see such attitudes as attributable to the agent, and ones he is properly accountable for. So the counterfactual specified by Levy is not a helpful (part of a) moral framework for evaluating morally agents, their implicit and explicit attitudes, and actions.

The case of individual B also helps us to see that it is a mistake to emphasise the lack of integration of implicit biases as a definitive feature of their characterisation. By identifying implicit biases as patchy endorsements - as states that, by virtue of their characterisation as 'patchy', are not well integrated with the agent's explicit attitudes - Levy marginalises non-conflict cases of implicit attitude. But this is a mistake, insofar as implicit biases still have an explanatory role in our understanding of the full scope of discriminatory behaviour in such cases. In the case of individual A (the racist), there will be some subtle automatic or autonomic responses that are not under the control of the agent's explicit attitudes (his fear responses, increased sweating and blink rates or dilated pupils, for example (Dovidio et al 1997)). In the case of individual B (the protocol-adhering racist) there will likewise be these aspects of his behaviour which, whilst supported by his explicit racist attitudes, are divergent from his explicit intention to adhere to protocol and avoid discrimination in hiring. Identifying these aspects of behaviour is crucial both for adequate moral evaluation, and for being able to target interventions to tackle discrimination appropriately.

In sum, it is unsatisfactory to characterise implicit attitudes as patchy endorsements which are necessarily poorly integrated with and divergent from the agent's explicit attitudes.[7] This characterisation omits, by conceptual fiat, cases of aligned implicit and explicit attitudes. And this omission is problematic insofar as it fails to provide adequate resources for moral evaluation, nor for complete prescriptions about interventions to prevent discriminatory actions.

(ii) In-between beliefs. Schwitzgebel (2010) provides us with a different model for thinking about implicit biases: as cases of in-between belief. The idea behind this picture is that, on a broad track dispositional model of belief,[8] there will be cases in which it is most appropriate to say that we 'kind of' believe some proposition; our cognitive position is pretty much in-between. As Schwitzgebel illustrates, we suppose an individual has some of the dispositions associated with holding some belief, P: believing that family life is more important than work. She might avow P,

7 Indeed, we could think that biases can be aligned whilst not being well-integrated, in the sense of not standing in the appropriate inferential or reasons-responsive relations. But this distinction is elided in Levy's claims.

8 Schwitzgebel notes that the idea of in-between belief can also be rendered cogent on other models of belief, functionalist, representationalist, and so on.

act as if P in context C1 (for example, prioritising school pick up). But the agent also has other dispositions that are indicative of holding the belief not-P (that it is not the case that family life is more important than work). For instance, the agent might act in context C2 as if not-P were true (working late, drifting into work thoughts whilst on family holiday, and so on). The idea is that in such cases, it is inappropriate to suppose that the agent flat out believes P or not-P; nor to ascribe both beliefs; nor to posit that she switches back and forth between P and not-P. Rather, she *kind of*, or in-between believes, both.

Implicit biases can be understood on this model, Schwitzgebel suggests. (It is worth noting that he offers implicit biases as a case in support of his claims about in-between belief, rather than as claims about a theory of implicit biases. Nonetheless, we can still ask whether this model really is a useful way of thinking about implicit biases.) An individual, such as C from our test cases, has some dispositions that we stereotypically associate with egalitarian commitments to fair treatment and anti-discrimination, and beliefs in racial equality. She may avow such beliefs, ensure her workplace has policies to address discrimination, attend anti-racism activist events. But an implicitly biased individual may also have and manifest dispositions associated with the racist belief that differently racialised individuals are not equal; she may react with greater hostility in interracial interactions or express surprise when discovering the academic excellence of one of her black students. These are dispositions stereotypically associated with racist beliefs or attitudes. In such cases we say that the agent in-between believes in racial equality. Some of her dispositions are consistent with this belief; others not - so we should not ascribe the belief wholesale, nor its contrary to her; rather, she *kind of,* or in-between believes, in racial equality.

This model of in-between belief seems helpful when we consider conflict cases of implicit biases - cases where the biases are in tension with the agent's explicit beliefs and values, or thwart her pursuit of them. And, it does so in a way that seems to satisfy some of our desiderata - we find implicit biases unpacked in terms that cohere with extant posits in our theories of cognition (beliefs, understood in dispositional terms). This provides us with a framework in which we can start to think about moral evaluation - to the extent that we have views about responsibility for belief, we can consider importing those to apply to these cases of in-between belief (to what extent are the agent's dispositions attributable to her, reflective of her evaluative stance, and so on). We see this aspect of the view in Schwitzgebel's remarks about responsibility for implicit bias (he argues individuals are responsible insofar as implicit biases are attributable to the agent, and reflective of her personality and character).[9]

However, this analysis is not so helpful when it comes to understanding the role of implicit bias in non-conflict cases - cases where some of the agent's behavioural dispositions (those related to implicit biases) line up with the agent's explicit attitudes and beliefs (in particular in case A above). In such cases, there is clearly no issue of 'in-between belief", since the kind of behavioural dispositions (due to implicit cognition) that we find line up pretty neatly with the agent's explicit beliefs or attitudes. Whilst not strictly speaking incompatible with recognising the explanatory significance of implicit biases, adopting in-between belief as a model of implicit bias obscures the role implicit bias can play in cases where the implicit and explicit attitudes are aligned. This is especially so since, according to Schwitzgebel, we ascribe beliefs based on the observed dispositional profile of the agent. And it is the dispositions 'features of which are broadly recognisable to normal attitude ascribers' (2010, p.12) that determine whether a belief or attitude is ascribed. Consider how this goes in conflict cases: we find that the agent behaves 'as if' they had racist beliefs despite their explicit avowals, and the behavioural outcome itself leads us to ascribe the in-between belief. But in non-conflict cases, such as that of individual A, we find the behavioural outcome is explainable by the agent's intentional decision to discriminate. The problem with this move is that the various dispositions associated with harbouring implicit race biases include also subtle differences in interpersonal interactions, autonomic and automatic responses, as described above. Yet insofar these behavioural dispositions are not part of the repertoire of 'normal

9 http://www.newappsblog.com/2015/03/on-being-blameworthy-for-unwelcome-thoughts-reactions-and-biases.html

attitude ascribers', they may be overlooked when there are intentional discriminatory decisions that explain the agent's behaviour. Accordingly, it is not obvious that the dispositions that manifest in, for example, unintentional micro-behaviours will be given due role in explaining behaviour in cases of aligned implicit and explicit attitudes. Whilst some of the fine-grained behavioural dispositions (the autonomic responses) are well captured in conflict cases, as 'in-between belief'; they are not aptly described as such in these non-conflict cases. This obscures the role of the dispositional profile associated with the implicit aspects of cognition.

I don't want to overstate the case against the in-between belief model of implicit attitudes. The claim isn't that there are no resources to make sense of the sort of non-conflict cases I have described. Rather the worry is that the role of implicit bias in such cases is easily obfuscated or obscured on this account, and we should be alert to these difficulties in order to deliver a more satisfactory model.

(iii) Unconscious beliefs. The last two views posited revisionary states to explain implicit biases: patchy-endorsements and in-between beliefs. On Mandelbaum's view, such revisionary states are not needed: rather implicit biases are best understood as 'honest-to-god propositionally structured mental representations that we bear the belief relation to' (2015, p.7). Unlike the sort of beliefs we are familiar with, though, implicit biases are unconscious beliefs. This means that (not necessarily, but at least typically) they do not figure in our conscious cognitions, but can function inferentially, and in reasons-responsive ways, beyond the reach of our reflective awareness. On this view, whilst a person such as individual C may have an explicit belief, for example, that *race is irrelevant to the quality and suitability of an applicant*, she may nonetheless harbour an unconscious racist belief, with the propositional content *black people are less likely to be suitable candidates*.

This view is primarily motivated by a swathe of research findings that are difficult to make sense of on the assumption that implicit cognition is purely associative (the sort of findings that also inform Levy's commitment to the claim that implicit attitudes may have propositional content, albeit (for Levy) content that is inferentially sensitive in a patchy sort of way). For example, Mandelbaum canvases studies in which implicit cognition seems to respect the dictum 'the enemy of my enemy is my friend' - difficult to make sense of on a purely associative model, where one might expect negative associations to accrue to the enemy of my enemy (two relations of dislike being at issue). Instead positive associations with the target object 'the enemy of the enemy' are recorded, which suggests sensitivity to content (Gawronski et al 2005).

Accordingly, this view stakes a claim to meeting the first desiderata of empirical validity. Moreover, as emphasised by Mandelbaum, this model crucially helps us to predict the sorts of strategies worth evaluating for efficacy in mitigating implicit bias. Those which focus only on the counter-conditioning of associative states will fail to exploit the extent to which implicit biases may be sensitive to inferential interventions. Likewise, it is also pretty clear that the posits of this view cohere with those of our extant theories of cognition, so the unconscious belief model of implicit bias fares well with respect to this desiderata, at least at first blush. Finally, insofar as we have available to us frameworks for understanding the moral evaluation of beliefs and actions informed by them (see for example Hieronymi, 2008), this view appears to stand well with respect to the our desiderata regarding the framework for moral evaluation of implicit biases.

How does this view fare in dealing with our test cases, however? We might first worry whether the unconscious belief account will capture all aspects of the phenomenon described in our test cases, which include the activation of affective and autonomic responses as well as responses that may more aptly be described as belief. At best this account has to see these as the causal upshot of the unconscious beliefs. Even with that caveat, the difficult cases for Mandelbaum's view are those in which there is alignment between the implicit and explicit biases. How might the unconscious belief model handle these cases? One option would be to say that there are two beliefs here: one explicit (the racist belief), the other unconscious (with the same content). In case B (the protocol-adhering racist) the unconscious belief is causally efficacious even whilst the explicit one is not, in case (A) both beliefs are causally efficacious. This is problematic, insofar as it commit us

to the idea that beliefs are individuated, and distinguished, not by their content (or indeed functional role) but according to whether they are conscious or not (hence we have two beliefs here, rather than one). This contention sits in tension with the idea that a belief might remain the same (have the same content, play the same functional role) whilst being sometimes conscious, sometimes not. In response, one might insist that the unconscious and conscious beliefs have different content - and perhaps sometimes they do; but to suppose they *always* did would be ad hoc.[10]

Accordingly, another option might be to say that there is one belief (the racist belief) and that it is held both consciously and unconsciously. That is something we should want to avoid saying on pain of inconsistency. Alternatively, we might say that this is belief is sometimes held consciously, sometimes unconsciously. This is surely true of many of our beliefs: my belief that walls are impassable is unconscious until I explicitly call to mind the properties (solidity, etc) that the wall has – yet it is causally efficacious even when it remains unconscious (I do not attempt to traverse the wall). This kind of move would help the unconscious beliefs view make good sense of non-conflict cases. But this then seems to require that we treat cases of conflicting biases quite differently from cases of non-conflict. In conflict cases (such as that of the biased egalitarian) the thought is precisely that appeal to an agent's run-of-the-mill beliefs is insufficient to explain her behaviour – we need to consider other aspects of her cognition to explain what is going on. Mandelbaum suggests that we call that other aspect an 'unconscious belief', which conflicts with the agent's conscious beliefs. However, in the non-conflict cases, we just appeal to the agent's run-of-the-mill beliefs, which – like other beliefs – are sometimes conscious, sometimes not. The unconscious belief account then owes us an explanation of why the mechanism that produces the automatic and autonomic features of behaviour is different in conflict than in non-conflict cases.

On the other hand, if it turns out that these 'other aspects' that explain the automatic and autonomic features of her discriminatory behaviour are just beliefs (sometimes unconscious ones, in the way that my belief about walls is sometimes unconscious, sometimes conscious and occurrent) then we seem to lose some explanatory power that we were supposed to gain by positing supposedly distinctively unconscious beliefs in the first place. Why is it difficult for agents to be aware of these racist unconscious beliefs if they are just run of the mill beliefs? (I don't have this difficulty with my beliefs about the solidity of walls.) Perhaps there is an important distinction to be had between non-occurrent beliefs and unconscious beliefs – but we need to see what that distinction is and how it helps us to make sense of the role of implicit biases in non-conflict cases.

Before we see how those claims might be fleshed out, we are not in a position to evaluate fully whether or not the unconscious belief account can satisfactorily meet our desiderata – especially in terms of providing a framework for moral evaluation; and for providing testable hypotheses for effective interventions to mitigate implicit biases (for example, how similar to conscious or non-occurrent beliefs do unconscious beliefs have to be for the same framework for moral evaluation to apply?). Thinking about our test cases shows us that there is more work to be done to complete the unconscious belief model and evaluate whether it is able to meet our desiderata.

(iv) Aliefs. Gendler has recently introduced the notion of *alief* to capture certain aspects of our cognition that are different in kind from our beliefs and other propositional, reflective and rationally guided states. These other kinds of mental states, aliefs, are associative, automatic and arational. On Gendler's characterisation:

> A paradigmatic alief is a mental state with associatively linked content that is representational, affective and behavioral, and that is activated— consciously or nonconsciously—by features of the subject's internal or ambient environment. Aliefs may be

10 One might insist that since they have different functional roles, they are different mental states: but on familiar views whereby mental states are individuated by their functional role, this will be different to reconcile with Mandelbaum's insistence that the mental state at issue, albeit unconscious, remains a belief. Thanks to Komarine Romdenh-Romluc for this thought.

either occurrent or dispositional (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. Positing aliefs,Gendler suggests, can help us make sense of the cognitions that are at work where people display responses or dispositions that are at odds with their beliefs. I *believe* the Skywalk is perfectly safe; but my *alief* is 'High up! [Representation]; Scary! [Affective response]; Get off! [Readying of motor response]'.

We can readily think of implicit biases as a kind of alief, and indeed, the phenomenon of implicit bias is one that Gendler uses to elucidate the notion of alief (2008b p.553). Consider our example of the biased egalitarian (C): this individual has various egalitarian and anti-racist beliefs: that the race of applicants is not relevant to the quality of the application, that her interactions should not differ depending on the race of the interviewee. But her implicit bias, as manifested in greater hostility in her interactions, can be unpacked as a cluster of activated contents: 'black applicant [representational]; negative affect/fear [affective response]; hostile/avoid [readying of motor response]'. This is the agent's alief towards the black applicant, which is discordant with her other mental states, beliefs, that do not discriminate between the black and white applicants.

Gendler uses these discordant cases to elucidate the phenomenon of *alief* - cases in which an agent's 'salient occurent aliefs will come apart from her occurent reality-reflective attitudes' (2008b, p.554). But crucially for present purposes, we may also find aliefs that are norm-concordant: cases in which an agent's occurent aliefs are in accordance with her beliefs and other explicit attitudes (2008b, p.554). These 'norm-concordant' cases are what we have in non-conflict cases.

Accordingly, it looks like the alief model is well placed to help us to make sense of the test cases that have, thus far, posed difficulties for models of implicit bias. We can say that the racist (A) and protocol-adhering racist (B) have explicitly racist beliefs alongside the aliefs 'black applicant; negative affect; hostile'. That is, these agents have a cluster of associative, automatic responses that influence behaviour and judgement, alongside the role that their explicit mental states play. This looks like a viable model for dealing with our difficult test cases, then. With respect to our other desiderata, this model certainly departs from our existing theories of mind, and folk notions of psychological state - and it will certainly take some careful work to articulate how aliefs might fit into a framework for moral evaluation. If we take seriously Gendler's remarks that aliefs can be changed, by processes of habituation or counter-conditioning of cognitive responses (2008b, p.572-576), it looks like aliefs can provide a model that meets these desiderata. So, if the departure from folk psychology is well motivated there may be reason to accept it.

However, I do not think it is well motivated, primarily for reasons expounded by Greg Currie and Anna Ichino. They argue, first, that there is no reason to suppose that there are mental states, aliefs, as such. They claim that the contents Gendler identifies as 'clustering' in a way that underwrites a mental state - alief - may be causally related, but there is no reason to suppose they constitute a 'substantial entity' or sui generis state. Indeed, secondly, they claim these contents are best understood as familiar other states: perception or representation; affect and behavioural response. Whilst Gendler emphasises the 'tightness' of the clustering, Currie and Ichino point out that she also accepts that the causal relations between the contents can be broken - by habituation, or practice, say (2012, p.790). This is well explained by a view that sees them as clusters of causally related distinct states, rather than as one sui generis state, alief. In fact, they argue, many other states are closely causally related (believing you are being charged by a bull, fearing it, readying to flee), but there is no theoretical gain from identifying such a cluster with some new distinctive mental state.

I find the case they make compelling, and do not see the motivation for gathering together the automatic, associative states involved in implicit cognition into a new mental state, *alief*. In fact, I think there is good reason to refrain from doing so, if this move gives the misleading impression

of homogeneity in the mental states we call 'aliefs'. I have argued elsewhere (Holroyd & Sweetman 2016) that there is considerable heterogeneity in the sorts of associations involved in implicit biases, and that recognising this is important for understanding the different strategies that may be needed to combat implicit bias (see desiderata 5). For example, it may be that some of the phenomena we call implicit bias are co-activated representations (such as those which encode stereotypes); others co-activated representation and affect; others yet co-activated representations, affect and behavioural response. We need not suppose that the same tripartite structure will be found in all instances of implicit bias. Accordingly, we should not endorse a model of implicit cognition that is unmotivated, and obscures important differences between different implicit biases.

Note that this line of enquiry has been fruitful, however: it points us towards an understanding of implicit bias as constituted by different combinations of co-activated contents: representational or affective or activated motor responses, which stand in close causal relations (this may be the sort of 'cousins' of alief that Currie and Ichino suggest in their 2012, p.797). Perhaps this minimalistic understanding is sufficient for modelling the phenomenon: further investigation will be required to ascertain whether it satisfactorily meets our desiderata. But we can see how at first glance, this 'minimal model' may find some empirical support. It coheres with understandings of our psychology which recognise associative processing - though may depart from folk psychological models that leave little room for this. Careful work will be needed to ascertain whether this sort of cognition falls within the remit of moral evaluation or warrants ascriptions of responsibility - but we may compare other sorts of automatic processing to make headway on this, so have a framework to work with. And this kind of model can accommodate the diversity of processes involved in implicit cognition - either representational, or affective, or behavioural - which may be important in identifying the full range of phenomena that require intervention. However, before it is worth pursing such a model further, we should consider a challenge to the idea that we need a model of implicit cognition at all.

(v) Traits. Perhaps our attempts to model implicit cognition are misguided, because in doing so we suppose that there is such a thing as implicit biases or implicit attitudes. But on Machery's view, this distinction is itself wrong-headed, at least in relation to implicit attitudes (2016). Machery's claim is that there is no distinction between implicit and explicit attitudes, since attitudes are traits, and so cannot manifest the property of being either implicit or explicit. The idea here is that, if we take seriously the idea that attitudes are dispositional profiles, then we reject the claim that they are mental states. Attitudes are dispositions to cognize, respond affectively, and behave in certain ways. On this view, then, attitudes cannot be implicit or explicit, since they are traits, and traits are not the sort of things that can be implicit or explicit.

What do we say, then, about those sorts of things we might have called 'explicit attitudes': the attitudes of the biased egalitarian, for example, that seem to be at issue when she proclaims that she affirms the value of racial equality? On Machery's view, these are judgements about attitudes, rather than expressions of attitudes themselves - judgements that we have certain attitudes. And, these judgements can be more or less accurate: the biased egalitarian inaccurately thinks she is non-discriminatory, but parts of the trait she is reporting on are not accurately detected; those dispositions to discriminate in some contexts. Hence the departure between her report and the dispositional profile that we observe when she discriminates.

Where does this leave the implicit/explicit distinction, such that we can make good sense of our test cases above? Whilst I am generally sympathetic to the trait view of attitudes, it does not yet provide the resources to help us make good sense of implicit biases per se. This analysis just pushes the distinction back a level - attitudes themselves are not implicit; but components of attitudes may be. On this view, implicit biases will be a component of the trait (along with various other things). Some of the components (the psychological basis) of the trait may include the sorts of automatic responses identified as part of implicit cognition, as Machery seems to acknowledge. So far as the trait view offers us an theory of attitudes, it tells us that these attitudes are not, themselves, implicit cognitions. But it does not give the rest of the story about implicit cognition. And insofar as there

are implicit biases or associations that constitute the psychological basis of attitudes we want to know what it is about *these* states that characterises them as implicit and delivers a model that meets our desiderata.

IV

Preliminary Conclusions: I have suggested that five ways of thinking about implicit biases - as patchy endorsements, as in-between beliefs, as unconscious beliefs, as aliefs, or as part of the psychological basis of attitudes - are not yet satisfactory in helping us to meet our desiderata for an account of implicit cognition. They are unmotivated, or leave no room for, or obfuscate, or are simply yet unable to speak to, the explanatory significance of implicit biases in what I have called non-conflict cases - cases in which the agents implicit and explicit racial attitudes are aligned.

I have suggested that what I referred to as the 'minimal model' above may fare better: this account sees implicit biases simply as causally related, or co-activated representational contents, or affective and behavioural responses. Some or all of these contents or responses may be involved in different cases in which implicit bias is operative. It seems to me fruitful to see what work the minimal model can do. There are two key thoughts that we should bear in mind when developing the minimal model. First, we should note that nothing in the minimal model entails that implicit biases are necessarily unconscious or impossible to be aware of. And I think there are good reasons avoid any such commitments. Recent empirical evidence suggests we may be able to gain awareness of the presence of implicit biases, or of their influence on behaviour. It might be difficult to have introspective access to biases, or to observe their effects in our cognitions and behaviours; but it is not clear that it is impossible to garner such awareness (see e.g. Pronin et al 2007, Hahn et al 2014).[11] That there may be variation in the extent to which awareness is possible in relation to implicit bias coheres with the minimal model, since we can draw on a host of explanatory resources to see why we might sometimes be alert to the activation of representational content or affective responses, and at other times unaware. This opens up the further question of in what sense this sort of cognition is 'implicit'.

Secondly, we might explore, in relation to the minimal model, the various kinds of control that we have with respect to implicit biases, and whether these distinguish implicit from explicit cognitions. Whilst implicit biases are not beyond our control,[12] it seems to be the case that the kind of control we may have in relation to implicit biases differs significantly from the kind of control we can exercise over other states that are guided by reflective deliberation: a framework focused on distinctive kinds of control may be of use in developing an account of implicit cognition. And we now have articulated some desiderata, and some heretofore ignored test cases, that can assist us in developing a more satisfactory model.

Department of Philosophy
The University of Sheffield
45 Victoria Street
S3 7QB
j.d.holroyd@sheffield.ac.uk

Thanks to audiences at Institutet för Framtidsstudier, Stockholm, the MAP conferences at KCL and Leeds, and to the CogSci research network at the University of Sheffield; each provided

11 I have discussed the different kinds of awareness we might have in relation to implicit bias in Holroyd 2014. I also consider various explanations for why we might be unaware of bias, even if such awareness is possible: self-deception, or confabulation, for example.

12 I have elsewhere argued, with Dan Kelly, for a model of ecological control that helps us to make sense of the kind of control at issue (Holroyd & Kelly 2015). See also Scaife (ms.) for a model of implicit cognition that focuses on the kinds of control we have over implicit biases.

constructive discussion in which the ideas of the paper were developed. This research was supported by the Leverhulme Trust Research Project Grant, on 'Bias and Blame', and I have benefited from fruitful discussions with co-investigators Tom Stafford, Robin Scaife and Andreas Bunge. Special thanks to Anna Ichino and Federico Picinali, for invaluable discussions of earlier drafts.

References

Committee on Maximizing the Potential of Women in Academic Science and Engineering, National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2007: 'Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering'. National Academies Press, http://www.nap.edu/

Cudd, Ann 2005: 'How to Explain Oppression: Criteria of Adequacy for Normative Explanatory Theories' *Philosophy of the Social Sciences*, 35(1), pp.20-49

Currie, Greg and Anna Ichino 2012: 'Aliefs don't exist, though some of their relatives do'. *Analysis* 72 (4), pp.788-798.

Patricia G. Devine, E. Ashby Plant, David M. Amodio, Eddie Harmon-Jones, and Stephanie L. Vance, 2002: 'The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice'. *Journal of Personality and Social Psychology* 82, pp.835–48.

Devine, Patricia G. 1989: 'Stereotypes and Prejudice: their automatic and controlled components'. *Journal of Personality and Social Psychology* 56, pp.5-18.

Dovidio John F. & Samuel L. Gaertner 2000: 'Aversive Racism and Selection Decisions: 1989 and 1999'. *Psychological Science* 11, pp.319–323

Dovidio, John. F. Samuel, L. Gaertner, Kerry Kawakami, & Gordon Hodson 2002. 'Why can't we just get along? Interpersonal biases and interracial distrust' *Cultural Diversity and Ethnic Minority Psychology*, *8*(2), 88-102.

Dovidio, John F.; Kawakami, Kerry; Johnson, Craig; Johnson, Brenda; Howard, Adaiah. 1997: 'On the Nature of Prejudice: Automatic and Controlled Processes'. *Journal of Experimental Social Psychology* 33(5), pp.510–40

Dovidio John F., Kawakami Kerry, Gaertner Samuel L. 2002: 'Implicit and explicit prejudice and interracial interaction'. *Journal of Personalty and Social Psychology*; 82(1), pp. 62-8

Fischer, John Martin and Mark Ravizza 2000: *Responsibility and Control: A Theory of Moral Responsibility,* Cambridge University Press.

Gawronski, Bertram, Eva Walther, and Hartmut Blank 2005: 'Cognitive Consistency and the Formation of Interpersonal Attitudes: Cognitive Balance Affects the Encoding of Social Information'. *Journal of Experimental Social Psychology*, 41, pp.618–626.

Jack Glaser, Eric D. Knowles 2007: 'Implicit motivation to control prejudice'. *Journal of Experimental Social Psychology* 44, pp.164–172

Gendler, Tamar 2008a: 'Alief and Belief'. *Journal of Philosophy*, pp.634-663

-------- 2008b: 'Alief in Action (and Reaction)'. *Mind and Language* pp.552-585

Glasgow, Joshua 2016: 'Alienation and Responsibility'. In Michael Brownstein & Jennifer Saul (eds.) *Philosophy and Implicit Bias*. Oxford University Press

Hahn, Adam, Charles M. Judd, Holen K. Hirsh, and Irene V. Blair 2014: 'Awareness of implicit attitudes'. *Journal of Experimental Psychology: General* 143 (3), pp.1369.

Hieronymi, Pamela 2008: 'Responsibility for Believing'. *Synthese* 161(3), pp.357–373.

Holroyd, Jules 2014: 'Implicit Bias, Awareness and Imperfect Cognitions'. *Consciousness and cognition*, 33, pp.511-523

Holroyd, Jules & Daniel Kelly 2015: 'Implicit Bias, Character, and Control'. In Jonathan Webber and Alberto. Masala (eds.) *From Personality to Virtue*, Oxford: Oxford University Press.

Holroyd, Jules & Joseph Sweetman 2016: 'The Heterogeneity of Implicit Bias'. In Michael Brownstein & Jennifer Saul (eds.) *Philosophy and Implicit Bias*. Oxford University Press

Jost, John T., Laurie A. Rudman, Irene V. Blair, Dana R. Carney, Nilanjana Dasgupta, Jack Glaser, and Curtis D. Hardin 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, pp.39-69

Lai, Calvin K., Maddalena Marini, Steven A. Lehr, Carlo Cerruti, Jiyun-Elizabeth L. Shin, Jennifer A. Joy-Gaba, Arnold K. Ho Arnold K.; Teachman, Bethany A.; Wojcik, Sean P.; Koleva, Spassena P.; Frazier, Rebecca S.; Heiphetz, Larisa; Chen, Eva E.; Turner, Rhiannon N.; Haidt, Jonathan; Kesebir, Selin; Hawkins, Carlee Beth; Schaefer, Hillary S.; Rubichi, Sandro; Sartori, Giuseppe; Dial, Christopher M.; Sriram, N.; Banaji, Mahzarin R.; Nosek, Brian A. 2014: 'Reducing implicit racial preferences: I. A comparative investigation of 17 interventions'. *Journal of Experimental Psychology* 143, pp.1765-1785

Levy, Neil. forthcoming: 'Implicit Bias and Moral Responsibility: Probing the Data'. *Philosophy and Phenomenological Research*

--------- 2015: 'Neither fish nor fowl: Implicit attitudes as patchy endorsements'. *Noûs*, 49(4), pp.800-823

Lippert-Rasmussen, Kasper 2014: *Born Free and Equal? A Philosophical Enquiry into the Nature of Discrimination* Oxford University Press

Mandelbaum, Eric 2015: 'Attitude, Inference, Association: On the Propositional Structure of Implicit Bias'. *Noûs* doi: 10.1111/nous.12089

Machery, Edouard 2016: 'De-Freuding Implicit Attitudes'. In Michael Brownstein & Jennifer Saul (eds.) *Philosophy and Implicit Bias* Oxford University Press

McConnell, Allen R., and Jill M. Leibold 2001: 'Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes'. *Journal of experimental social psychology* 37(5), pp.435-442.

Pronin, Emily, and Matthew B. Kugler 2007: 'Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot'. *Journal of Experimental Social Psychology* 43 (4), pp.565-578.

Saul, Jennifer 2013: 'Unconscious Influences and Women in Philosophy'. In Fiona Jenkins & Katherine Hutchison (eds.) *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press pp.39-60

Schwitzgebel, Eric 2010: 'Acting Contrary to Our Professed Beliefs or The Gulf Between Occurrent Judgement and Dispositional Belief'. *Pacific Philosophical Quarterly* 91, pp.531–553

Washington, Natalia & Daniel Kelly, 2016: 'Who's Responsible for This? Moral Responsibility, Externalism and Knowledge about Implicit Bias'. In Michael Brownstein & Jennifer Saul (eds.) *Philosophy and Implicit Bias* Oxford University Press