1.3

The Heterogeneity of Implicit Bias

Jules Holroyd and Joseph Sweetman

The term 'implicit bias' has very swiftly been incorporated into philosophical discourse. Our aim in this chapter is to scrutinize the phenomena that fall under the rubric of implicit bias. The term is often used in a rather broad sense to capture a range of implicit social cognitions, and this is useful for some purposes. However, here we articulate some of the important functional differences between phenomena identified as instances of implicit bias. We caution against ignoring these differences, as it is likely they have considerable significance—not least for the sorts of normative recommendations being made concerning how to mitigate the bad effects of implicit bias.

1 The Disparate Phenomena called 'Implicit Bias'

Philosophical attention has galvanized around the notion of implicit bias in recent years. Roughly, studies show that individuals harbour many implicit associations between mental constructs, such as 'salt' and 'pepper', or 'white' and 'good'. Sometimes associations concerning stigmatized social groups influence a decision or action. An implicitly biased decision or action is one that expresses or embodies implicit features of cognition, which distort or influence that behaviour. For example, the implicit association between the race category 'white' and evaluative term 'good' can influence people to judge more positively a CV with a white-sounding name on it than the same CV with a black-sounding name (Dovidio and Gaertner, 2000). Philosophers have been particularly concerned with those implicit processes that influence behaviour in undesirable and often discriminatory ways. Some of the questions that philosophers have been interested in are: what are the ethical implications of acknowledging the influence of implicit bias on decision and action? What are the consequences for our understanding of agency, responsibility, and how we ought to act? What is

epistemologically problematic about the operation of implicit bias? What kinds of material changes in the world are needed to address and mitigate the likely operation of implicit bias? (See e.g. Kelly and Roedder, 2008; Machery, Faucher, and Kelly, 2010; Holroyd, 2012; Gendler, 2011; Haslanger, 2008.)

What, exactly, is implicit bias? Understanding this is an important step in sensibly addressing these other questions. Our concern is that broad characterizations of implicit bias have led to misleading generalizations, and normative recommendations that may either be counterproductive, or at least less useful than they could be—so we will argue.

Let us start by observing some of the ways in which the term has been characterized and used. In her influential paper 'Implicit bias, stereotype threat and women in philosophy', Jennifer Saul characterizes implicit bias as

> unconscious biases that affect the way we perceive, evaluate, or interact with people from the groups that our biases 'target'. (2013: 40)

This is a useful functional definition: implicit biases are whatever unconscious processes influence our perceptions, judgements and actions—in this context, in relation to social category members (women, blacks, gays, for example).[1] However, there is some evidence that suggests that implicit biases are not always 'unconscious'. It is contentious that the participants are unaware of the cognition that is being implicitly measured in tasks such as the IAT (De Houwer, 2006; Monteith and Voils, 1998). Work on the correction of implicit race bias specifically suggests that some awareness of implicit bias is possible, if not likely (Wegener and Petty, 1995). It would not be surprising, given the argument to follow, if there were variations in awareness of different implicit associations. The debate about awareness of implicit processes is interesting, but is not our focus here. (For a discussion, see De Houwer, 2006; Fazio and Olson, 2003, Hann et al., 2013; Holroyd, 2014; see also Gawronski, Hofmann, and Wilbur, 2006).

More importantly for the focus of this chapter, the functional definition we have started with here leaves open the matter of precisely what processes constitute implicit bias, and in particular whether we are dealing with a singular entity or a range of psychological tendencies.

A further concern is that this usage seems to permit ambiguous use of the notion of implicit bias: sometimes 'implicit bias' is used to refer to an *output* such as a biased decision or judgement (for example: '[i]t seems very likely, then, that philosophers will display implicit bias against women'; Saul, 2013: 43). It is also

[1] Saul's focus is principally on harmful implicit biases, but she notes that there are a range of idiosyncratic and unproblematic biases.

used in a way that appears to refer to a mental state or process. This can be seen in remarks concerning people 'hav[ing] implicit biases' (Saul, 2013: 55). We think this encourages the tendency to suppose that there is a unified process or state ('implicit bias') that produces a distorting influence on judgement and action (also referred to as 'implicit bias').[2]

Elsewhere, the term 'implicit bias' has been used even more expansively. In addressing the epistemological implications of implicit bias, Tamar Gendler (2011) discusses the phenomena of racial categorization, stereotype threat, and the cognitive depletion that subjects experience after interracial interactions, all under the rubric of 'implicit bias'. In this context, then, implicit bias is being used to pick out a range of social cognitions (and affective states), including but not limited to, unconscious activation and application of stereotypes (involving conscious feelings of anxiety/threat), automatic categorization (of things or people into groups which are perceived as sharing properties), and effortful activity (such as suppression of biased responses or stereotyping).

We contend that whilst in some contexts this kind of expansive understanding of implicit bias can be useful (Section 2), it also has significant limitations and tends to obscure important differences between implicit associations (Sections 3 and 4).

2 The Usefulness of an Expansive Concept

For three reasons, this broad usage makes considerable sense. First, the processes at issue in (for example) Gendler's treatment of the issue are all automatic, difficult to discern from introspection, difficult to bring under reflective control, and as a result not governed by the same norms of reasoning as are reflective states (such as occurrent beliefs and desires). It is useful to identify a set of processes that share these features, and contrast them with the kinds of reflective processes to which philosophers have typically attended. This enables attention to be drawn to the large range of mental activity not encompassed by a focus only

[2] Perhaps this conflation of output and cognitive content reflects the confusion over awareness mentioned above. While there is no evidence that people lack conscious awareness of the cognitive content measured by implicit measures, there is evidence to suggest that, under certain conditions, this content may impact other processes and behaviours outside of conscious awareness (Gawronski et al., 2006; Hann et al., 2013). As such, the question of whether implicit bias is unconscious depends on whether one is referring to content or output. More precisely, the evidence seems to suggest that it is the content that may, under some circumstance, be outside of conscious awareness. One of the key points we make in this chapter is that distinguishing differing content is important in understanding implicit processes—the way content (i.e. mental representations/associations) impacts on behaviour (output).

on reflective, deliberative cognition, and the importance of recognizing this range and its role in our mental lives.[3]

Secondly, functional definitions such as Saul's are helpful if one is concerned—as is important—with articulating the widespread *effects* of implicit biases, and the worries that arise in relation to these. For example, if we want to focus on and articulate the patterns of discrimination in which implicit biases might be implicated, then attending closely to the nature of the implicit cognitions themselves is not the priority task. If the priority is in articulating those effects rather than looking at the processes that produce those effects and ways of combating them, then it is reasonable to talk of implicit bias as simply whatever implicit processes produced those effects. Such a priority is important in gaining recognition of the pervasive nature of the problem. (One cannot encourage people to adopt strategies to combat the problem if they do not agree that there is a problem.)

Another consequential reason for subsuming a number of phenomena under the notion of implicit bias is that it makes it more likely that certain important claims about implicit bias are true. For example, Saul makes the general claim that 'human beings are strongly influenced by a range of disturbing and often unconscious biases' (2013: 40). And indeed, this claim is likely to be true if the notion of implicit bias is broadly construed to include a range of implicit social cognitions. The claim that 'we are all likely to be implicitly biased' (Saul, 2013: 55) will be true if 'implicitly biased' refers to a range of phenomena extending to a number of different negative and socially consequential implicit social cognitions.[4]

We do not mean to suggest that there is any sleight of hand here: gaining traction in addressing the effects of implicit bias requires garnering agreement on the claim that almost all of us will need to reconsider the ways in which our judgements and actions may be influenced—ways we would find surprising and perhaps uncomfortable. For these purposes, a broad characterization of implicit bias is legitimate and useful.

[3] One might argue that existing notions already perform these useful functions, such as those of automaticity or system 1 functioning. However, the notions of implicit social cognition, automaticity, and system 1 processes are all examples of dual-process models of the mind with the distinctions between these terms simply reflecting the particular area of cognitive psychology (i.e. memory, attention, and decision making, respectively) that gave rise to them. (For an excellent account of this history, see Gawronski and Payne, 2010.)

[4] That is: for a particular bias, b, it may not be probable that an individual has that bias—but for a range of biases, b1-bn, it is probable that an individual has a bias in that range. So it remains true that all individuals are likely to have some biases.

However, we think that there are two dangers in the philosophical discourse about implicit bias (which amplify each other). Firstly, we want to suggest that there are some dangers in regarding 'implicit bias' as a catch-all for a range of implicit (and not so implicit) social cognitions; doing so permits generalizations that may not be warranted. Secondly, we want to raise concerns about the tendency to overlook the ways in which processes that fall under the rubric of implicit bias may differ, either functionally or structurally, because attending to these differences probably has important implications for the normative recommendations made about how to combat problematic implicit biases.

This tendency is reinforced by the philosophical discourse concerning implicit bias, which speaks to 'the effects of implicit bias', 'the ethical implications of implicit bias', 'the epistemological implications of implicit bias', and so on. (For examples of such usage, see Gendler, 2011; Saul, 2013; Machery et al., 2010; Holroyd, 2012). This kind of discourse implies that the concern is with a certain homogeneous phenomenon (implicit bias) and its effects, and plays down the idea that there might be differences within the phenomena falling under the rubric of implicit bias.

It is our contention that, for some purposes, these differences matter. Perhaps most importantly, the differences matter to the kinds of normative recommendations needed concerning how to mitigate or remove the influence of implicit biases.

3 Implicit Processes and Different Kinds of Implicit Association

Philosophers are not alone in making assumptions about the unified nature of the phenomena. Amodio (2008) observes that 'researchers [including empirical psychologists] have generally assumed that implicit stereotyping and evaluation arise from the same underlying mechanism' (7). In the following sections we articulate the reasons to suppose that implicit bias is functionally heterogeneous, and that this heterogeneity matters considerably. We also consider the reasons for holding that the processes underpinning implicit biases are heterogeneous, but raise concerns about one dimension along which these processes have been distinguished.

In the literature from empirical psychology, we find reference to 'implicit processes' rather than implicit biases (Amodio and Mendoza, 2010; Nosek, Hawkins, and Frazier, 2012; for a review, see Gawronski and Payne, 2010). This suggests that what is at issue is a set of processes which share the property

of being 'implicit'—generally (and not uncontestedly[5]) under the radar of reflective introspection, difficult to bring under reflective control, and quick and efficient.

Amodio (in accordance with most other psychologists working in this domain, e.g. Fazio, 2007; Greenwald et al., 2002) understands these implicit processes essentially to consist in the utilization of 'associations stored in memory' (2010: 364). Associations[6] are discerned by tests such as the implicit association test (IAT) and affective priming task (Fazio et al., 1995). In the IAT (the most popular implicit measure) the implicit associations present in an individual's cognitive structures are revealed in the swiftness of response in categorizing concepts into pairs. A range of experimental tests aims to reveal individuals' associations, and thereby identify factors that may play a role in perception, judgement, and action, but which go unreported in reflective (explicit) statements of what guided behaviour (either because of self-presentation worries; Fazio and Olson, 2003) or simply because such associations are not readily detectable by the agent (see Brownstein and Saul, this volume, for a description of implicit measures such as the IAT).

What kinds of associations are at issue here? A number of central cases of implicit bias have concerned philosophers (there are numerous studies, but these have received significant attention). The associations between social category and stereotypic or negative notions are cause for particular concern. Such associations have variously been found to guide the evaluation of CVs, produce shooter bias, and affect interracial interactions (see e.g. Saul, 2013; Kelly and Roedder, 2008; Machery et al., 2010). Are the same associative mechanisms, or kinds of mechanism, involved in each of these cases? Do the associations all function in the same way, or are there important differences?

We advance two claims here: first, that evidence indicates that different associations have different characteristics. Accordingly, there is reason to doubt that all generalizations about implicit bias can be substantiated, and to be

[5] See e.g. Monteith, Voils, and Ashburn-Nardo (2001); De Houwer (2006); Fazio and Olson (2003); Nosek et al. (2012).

[6] The idea that memory, and the mind more broadly, is associative is a long-held view in philosophy. Subsequently, psychologists and neuroscientists concerned with learning and memory have adopted this idea. However, there is work to suggest that this associative picture of mind may be fundamentally flawed (see Gallistel, 2008; Gallistel and King, 2009; Gallistel and Matzel, 2013). While discussion of this point is outside the scope of the present chapter, we feel it important to acknowledge and to make clear that our argument is not fundamentally based on an associative picture of the mind. Regardless of the way memory is organized and instantiated in the brain, we believe that it is a mistake to suppose that the cognitive processes at work in implicit biases are all relevantly similar.

cautious about talk of 'implicit bias' simpliciter. Secondly, drawing on the work of Amodio (2008) and Amodio and Devine (2006), we consider whether it is appropriate to understand that one dimension along which implicit bias differs, both in terms of content and underpinning structure, is in terms of whether implicit associations are semantic or affective. These terms refer to ways of categorizing kinds of associative process: firstly, according to whether the content of the association concerns the meaning of the associated constructs, or the positive or negative affect accompanying a construct; and secondly, in terms of the processes underpinning these different contents. We argue that this distinction is problematically deployed in their empirical studies, and so whilst we might endorse the general claim that implicit associations differ functionally, in respect of how they seem to operate in relation to other beliefs and behaviours, we do not endorse the more specific claim that these differences are captured or predicted by either contents, or an underlying structure, that is understood in terms of the semantic/affective distinction. However, if the claim about functional difference is right, there will be important consequences for making generalizations about implicit bias. In particular, there are implications for the general recommendations made concerning how to mitigate implicit bias, which we address in the Section 4.

3.1 Biases behaving differently: two kinds of heterogeneity

At first glance, it seems clear that the studies that have been focused upon involve different associations. Some studies test for gendered associations, others test for associations with racial categories, others still for age, sexuality, and religious or ethnic groups, and respective associations.[7] Clearly, there are different associations involved in the studies reported on. The strength of implicit associations between the following categories (inter alia) have been tested:

- Gender (gendered words: she, woman/he, man) and words associated with leadership (manager, director/worker, assistant) (Webb, Sheeran, and Pepper, 2010).
- Sexuality (images of gay and heterosexual kisses) and positive and negative words (Payne, Cooley, Loersch, and Lei, ms.).

[7] Other implicit associations that have little directly to do with social identity have also been studied, such as associations concerned with health and other foods, with objects of fear (such as snakes), and so on. Whilst these associations have garnered less interest from philosophers, they have been important in advancing understanding of the kind of cognitive processes at work in these implicit attitudes and in developing clinical interventions for addiction and various psychopathologies.

- Race (black and white name primes) and personal preferences (like/dislike) (Olson and Fazio, 2004); black and white faces and positive and negative words (Amodio and Devine, 2006).
- Gendered pronouns (he/she) and job titles (nurse, mechanic) (Banaji and Hardin, 1996).
- Ethnic/religious group words (Muslim/Scottish) and words associated with terror, or peace (Webb, Sheeran, and Pepper, 2010).

It is obvious that there are different associations involved here; the relevant associations hold between relata with different content. But even this obvious fact is obscured by talk of 'implicit bias' simpliciter, without note of which associations are in play. Attending to this prompts us to ask what these differences in content amount to; should we expect all of these associations to behave in similar ways? We claim that there is reason to suppose not, and that this has implications for philosophical discussion about, and practical recommendations relating to, implicit biases. One explanation for the functional differences we outline in Section 3.2 appeals to the difference in contents. This means that it would be of utmost importance to attend to the content of implicit biases, rather than talk about implicit bias in generalized terms, if one is concerned to outline the effects, and ways of combating discriminatory behaviours.

A second kind of explanation might appeal to different processes involved in implicit cognition. For example, some have argued that there is reason to suppose that the various implicit measures are accessing discreet and non-unified implicit associations (such that they do not all cluster together to form an 'implicit attitude'), or perhaps different kinds of implicit processes. The IAT is one amongst a number of implicit measures; that is, tests which attempt to 'get at' individuals' implicit associations (such as their implicit race associations). A number of authors have pointed out that individuals' scores across implicit measures weakly correlate (that is, showing an implicit association on one measure does not correlate with showing an implicit association on another measure). For example, Fazio and Olson (2003) cite various studies in which they observe the 'disappointing correlations among various implicit measures', and report that 'in our own lab we have repeatedly failed to observe correlations between IAT measures and priming measures of racial attitudes' (277).

In a survey article, Nosek et al. (2007) argue that one of the best explanations for this weak correlation is simply the range of processes being tested for by the various implicit measures:

> [t]he relations may also reflect heterogeneity of cognitive processes that contribute to the various measures. The term *implicit* has become widely applied to measurement methods

for which subjects may be unaware of what is being measured, unaware of how it is being measured, or unable to control their performance on the measure. Identification of the cognitive processes that contribute to different measures will promote a more nuanced description and categorization of methods based on the particular processes that they engage. (277)

The idea for our present purposes is that the lack of correlation between implicit measures, as Nosek claims, is explained by the different processes or cognitive structures that each measure is tapping into. Nosek et al. suggest:

> The next generation of research in implicit cognition will likely revise the simple implicit–explicit distinction and introduce a more refined taxonomy that better reflects the heterogeneity of cognitive processes that are collectively termed *implicit*. (267)

Accordingly, there are two ways in which implicit biases might be heterogeneous. Firstly, we may observe functional heterogeneity in the way that different implicit associations operate (perhaps explainable by differences in content); and secondly, there may be heterogeneity in the processes underpinning different implicit associations. In this chapter we remain agnostic as to whether the heterogeneity manifest in implicit associations is attributable to different content or to different underlying processes. (Whilst the content explanation could do the explanatory work, it is possible that better formulated understandings of the structural differences might also have explanatory power.) Our main contention is that philosophers also need to be alert to the dimensions of heterogeneity in the ways that implicit biases operate, and possible distinctions between different kinds of implicit cognitive processes, that might explain this, for two reasons. Firstly, because evidence supporting such a taxonomy is relevant to precisely what generalizations can be made about implicit associations. Secondly, because the way that the distinctions are drawn may themselves require philosophical scrutiny. We return to this point in Section 3.2.

Regarding the first concern, about the generalizations that are warranted, one illustration of this pertains to the claims that philosophers have variously made about individuals being afflicted by implicit bias *irrespective* of their explicit beliefs. But there is reason to suppose that this generalization cannot be made. With respect to some associations, this claim seems true: in tests for implicit associations between gendered pronouns (he/she) and stereotypical roles (nurse/secretary), Banaji and Hardin (1996) found no difference in the extent of implicit biases between individuals who, on self-report measures, scored either high or low in sexist beliefs (139). In contrast, in studies reported in Devine (2002), it appears that individuals who held non-prejudiced behaviour to be important in itself display less race bias on race IATs (which require pairing black and white

face or names with positive or negative words); that is, the explicit beliefs and attitudes an individual held did seem to correlate with the degree of implicit bias they manifested. (See also Nosek et al., 2007 (277–8) for discussion of the cases in which self-report measures seem to correlate with implicit attitudes.)

Crucially, the heterogeneity of implicit biases in this respect means that some generalizations about the relationship of implicit associations to explicit beliefs—such as that implicit biases are independent of explicit attitudes—cannot be substantiated.

We have in this section distinguished between two ways in which implicit biases might be heterogeneous: functionally, or in terms of the underlying processes. We provided some evidence that supports functional heterogeneity, and indicated the reasons for which some have suggested that there may be heterogeneity in terms of the underlying processes involved. Of course, one explanation for these bits of evidence could be simply that the experimental designs did not always produce or measure the effects that they should or could have (cf. Nosek et al., 2007: 276). Nonetheless, the findings should give reason to exercise caution about claims that are general in nature, and that make recommendations for the regulation of bias that suppose general applicability of such recommendations. In Sections 3.2.a–3.2.c we provide further considerations in support of the claim that the best explanation of this heterogeneity is not experimental deficit, but rather differences between implicit associations and their operation.

3.2 Distinct associations with distinct behavioural influence

We have identified associations that are obviously, on the face of it, different. We have noted that some of these biases appear to stand in different relationships to explicit beliefs. This suffices for our central message of caution regarding the generalizations that can be made about implicit biases (a message we shall elaborate in Section 4). At least in this respect, then, generalizations about implicit biases are mistaken. This is significant, as there is a tendency to suppose that implicit biases are unrelated to explicit beliefs, and this may have further implications for how questions such as control, responsibility, and accountability are considered.

In this section we consider a further way in which implicit associations may differ; namely, with respect to the influence they exert on different kinds of behaviour. This dimension of functional heterogeneity has been articulated in the context of empirical studies that aim to differentiate between different underlying processes: 'semantic' and 'affective' associations. If we take these experimental results at face value, then there would be reason to suppose that it

identifies some underlying structure to the heterogeneous implicit processes. However, we argue that there are reasons to worry about this distinction, and that it should not, as presently articulated, be endorsed. This does not, however, undermine our central claim that implicit biases differ in important ways; there seems to be some important functional heterogeneity, though it is not best captured in terms of semantic and affective associations. Moreover, our discussion reinforces the claim that philosophers should attend to the ways in which psychologists are distinguishing different implicit biases. We explore the implications of this claim in the final section.

We start by presenting the further dimension of functional heterogeneity, then explain—and critique—the conceptual framework used to articulate this in terms of the heterogeneous underlying structures, by empirical psychologists.

3.2.1 DISTINCT ASSOCIATIONS

Amodio and Devine (2006) attempted to isolate the operation of different associations, and test for the presence of each. In order to do this, they constructed two race IATs. One was designed to test for associations between race and certain stereotypic traits: white/black, and mental (e.g. brainy, smart, educated) or physical (athletic, agile, rhythmic) constructs. They supposed that individuals might hold these implicit associations (such as a stronger association between *black* and physical constructs and between *white* and mental constructs) without also having negative attitudes or affect associated with that racial category (in common and imprecise parlance, an individual might hold a stereotype without having negative attitudes or disliking the stereotyped individuals). The second IAT was designed to test for these latter, negative affect-laden associations by asking participants to pair black or white faces with pleasant or unpleasant constructs (respectively: *love, loyal, freedom*; *abuse, bomb, sickness*).[8]

The striking—and crucial for our purposes—finding was this: 'the participant's scores on the two IAT's were uncorrelated' (Amodio and Devine, 2006: 14). That is to say, the extent to which individuals expressed the mental/physical associations *was not correlated* with scores on the second IAT for negative implicit attitudes.

Why is this significant? Firstly, it suggests that the two associations were in some subjects operating independently (Amodio and Devine, 2006: 655). Whilst we might expect many implicit associations to go in step (for example, we might expect an individual who implicitly associates black men with danger to also have

[8] This kind of attitude (stereotyping without negative affect) is termed 'benign racism' in analyses of racism. See e.g. Garcia (1996).

implicit negative affect—fear—towards them), this study indicates that at least some implicit associations about the same group are held independently. (Note that this observation is more readily made once we attend carefully to the different contents of implicit associations, obscured by generalized speak of 'implicit bias'.)

Secondly, these findings indicate that there may be variation across individuals with respect to which associations are operative in producing implicitly biased perceptions, judgements of or actions towards a particular group. For any implicit association, some individuals may have it and others may not (which is consistent with the variation in affective association found in studies by Devine et al., 2002). But the presence of (e.g.) one kind of implicit race association does not entail the presence of other forms of implicit race associations[9]—and conversely, the absence of one implicit association does not entail that one is free from other problematic implicit race biases.

Even if much of the time, or in many subjects, implicit associations work in concert, if there are distinct associations then it will be important to understand further the ways in which they may differ. This is of crucial import, given the differential behavioural outputs that these two implicit associations correlated with, which we now describe.

3.2.2 DISTINCT INFLUENCES ON BEHAVIOUR

Not only did the studies indicate that different implicit associations were not correlated; they also indicated that the different associations uniquely predict different behavioural outcomes. In the study by Amodio and Devine (2006), participants were asked to make judgements about the competences of a potential test partner, and then asked to sit and wait for their test partner to enter the room. The—in fact, fictive—test partner was indicated to be African American. Seating distance was measured as a behavioural indicator of positive or negative affect. Experimental participants who displayed strong associations on the race IAT for the mental/physical constructs, described in Section 3.2.a, made judgements about the competence of their test partner consistent with stereotypes (such as competence on questions about sports and popular culture, rather than mathematics). But these kinds of association did not predict greater seating distance from the test partner. On the other hand, manifestation of strong negative evaluative associations on the affect-based IAT uniquely predicted

[9] As Alex Madva has pointed out (correspondence), it might permit us to infer the increased probability of other sorts of implicit race association, even if the correlation is low.

seating distance (greater negative associations correlated with greater seating distance), but not judgements of competence.[10]

So, one association seems to be implicated in the judgements and evaluations individuals made, the other in approach or avoidance behaviours. This provides further support for the worry we raise: that certain generalizations about predicted behaviours cannot be made across various implicit associations. Note that this is not at all surprising when we consider associations that differ in their target: we would not expect gender associations to predict behavioural outcomes regarding racial interactions. What is noteworthy here is that different race associations (that is, associations that concern the same target social identity) are operating independently and with different behavioural predictions.[11]

If different implicit associations seem to exert influence on different kinds of behaviour, then understanding this will be important in formulating strategies that aim to combat implicit bias. In relation to the particular associations at issue here, for example, if one is involved in a task such as evaluation of an individual's competence or intelligence, then mitigating the associations between race and mental or physical constructs that may influence that judgement will be of particular importance. On the other hand, if one is concerned with increasing the amount and quality of intergroup contact, one might focus on limiting or changing negative affective associations. We return to this point in Section 4.1.

3.2.3 A DIMENSION OF HETEROGENEITY: SEMANTIC AND AFFECTIVE ASSOCIATIONS?

The experimental results we have just presented support our thesis that implicit associations are functionally heterogeneous and may not readily admit of the sorts of generalization that have been made (concerning behavioural predictions and their relation to explicit beliefs, for example). However, these results are framed in empirical psychology in terms of two different kinds of association: semantic and affective. The idea is that this identifies a systematic difference in content, which is underpinned by a structural heterogeneity (along which the

[10] It is worth noting the study by Macrae et al. (1994), which seems to indicate that stereotypes can affect seating distance. Participants in whom stereotypes were activated sat further from the stigmatized individual. Is this finding in tension with that by Amodio and Devine (2006)? We think not. It is important to observe that the stereotype at work in the study by Macrae et al. was that of 'skinhead', which is likely to involve various associations (fear, hostility, aggression) that are more similar to the negative evaluative associations found to predict greater seating distance in the study by Amodio and Devine. Moreover, this finding drives home our overall point that it is difficult to make generalizations across different kinds of association. Attention must be paid to how different contents may produce different behavioural predictions and outcomes.

[11] Thanks to Alex Madva for emphasizing the importance of this point.

functional heterogeneity may be explained). The mental/physical constructs are identified as 'semantic' associations. Other examples of this kind of association are salt/pepper and woman/she. The unpleasant/pleasant constructs (used on the second IAT described in Section 3.2.a) are identified as 'affective' associations—those that have an affective valence. Other associations put into this category include those for which one relata is evaluative, either generally ('good/bad') or in more specific ways ('attractive/disgusting'). Generalizations about these two kinds of association, concerning their influence on behaviour, how they might be learned or unlearned, are then made. Here we have not adopted this way of conceptualizing the distinction, nor supposed that the differences described in the previous sections are underpinned by such a distinction—either in content, or in underpinning processing—and are reluctant to do so for the following reasons.

Firstly, even if we endorse heterogeneity in content, considerations of parsimony counsel against explaining these differences in terms of different underlying associative processes. That implicit associations dissociate, and generate different behavioural predictions, could be explained in terms of the content of the associations, without recourse to distinct underlying mechanisms.

Secondly, however, even at the level of contents, the distinction posited is itself problematic. Whilst it is coherent to draw such a distinction (between those associations which have affective content and those which do not), the way this distinction is deployed is problematic. For one thing, it seems inappropriate to characterize one side of the distinction as 'semantic'. How should we best make sense of the idea of a 'semantic' association? This category has been used to identify associations that hold between 'semantically related concepts' (Amodio, 2008: 8). But that idea seems deployed problematically in the study described in Sections 3.2.a and 3.2.b: the stereotypic association is not adequately characterized as a matter of the semantic meaning of *black* or white; nothing in the *meaning* of these terms is associated with mental or physical constructs (in contrast, the meaning of *woman/she* clearly is semantically related; a paradigm case of semantic relationship is between 'bachelor' and 'unmarried man'). The characterization might aim to pick out the fact that certain semantic content has become associated with the racial category, such that the two are associated in mind. But this does not help us to pick out one side of the distinction, as paradigm relata of affective associations (*good*, *attractive*, and so on) have semantic content which comes to be associated with one social group.

Perhaps what is at issue is the contents of a *schema* (or stereotype) for different racial categories (and other aspects of social identity) (Haslanger, 2008). Schemas are characterized by Haslanger as 'a patterned set of dispositions in response to one's circumstances' (212). Might we understand semantic associations in terms

of the contents of a schema, saying that if included in a schema, an association (a kind of pattern of thought) is semantically associated? The problem is that it is not at all clear that schemas do not include the sorts of association that have been classified as 'affective', as dispositions to respond could just as well be underpinned by affect as by cognitive understandings. (Haslanger is here drawing on Valian (2005), who denies that schemas have affective content. Valian writes that on her account, schemas are 'cold'. Her account 'is purely cognitive rather than emotional or motivational' (198). We believe our point to show that Valian's understanding of schemas, which is narrower than Haslanger's, to be mistaken in excluding affective content, if that is to include the negative affect that attaches to evaluative terms such as 'good', 'bad', 'loyal', 'evil', and so on.)

All this raises worries for characterizing one side of the distinction as 'semantic'. But the difficulty is not simply that this way of describing the distinction seems inapt; rather, the distinction does not seem to cut where it needs to in order for Amodio and Devine to draw their conclusions about the heterogeneity of content, or underlying processes (as affective or semantic). If the distinction is supposed to be between associations with semantic content and those without, then this distinction was not adequately captured by their experimental design, because some associations that are supposed to be on the affective side have semantic content (good, disgusting, and so on).

Perhaps the distinction captured by their experimental design is supposed to be between associations that are affectively valenced (with positive or negative affective 'pull') and those that are not. Some who endorse the primacy-of-affect thesis, according to which all concepts held have *some* valence, might worry about this characterization: everything, it seems, would fall into the 'affective' category. One might reject that worry: perhaps the 'affective' associations can be identified as those that produce affect above a certain threshold. Even still, whilst this may present us with a conceptually coherent way of drawing the distinction that the categories do not seem to be exclusive seems to pose difficulties for the thesis that there are two distinct kinds of content, which operate on two different underlying structures, about which generalizations and predictions can be made. This is especially so because the experimental studies utilize notions which incorporate both semantic content and affect (good/bad, attractive/disgusting, and so on).

Even those terms that are supposed to indicate semantic associations in Amodio and Devine's studies (intelligence, athleticism, smart) have both evaluative and semantic content (the characteristics are positive, good, features). So, we might at this stage claim that such a distinction is coherent (if not best described as 'semantic' and 'affective'), but that the studies by Amodio and

Devine do not exclusively track this distinction, insofar as the supposedly semantic associations had affective content, and the supposedly affective associations had semantic content. Given this, their deployment of the distinction cannot be used to support the claim that there are two kinds of association that generate distinct behavioural predictions.

An analogy can help us to make this point. Suppose one wanted to evaluate children's well-being, and one supposed that a dimension that might explain different levels of well-being is whether a child has an active father in their life, or is raised by a single parent. One might construct a study to evaluate this hypothesis. But that distinction on which the hypothesis rests is deployed problematically, and cannot be explanatorily useful in predicting different outcomes, because (obviously) some single parents are fathers. In order to test the hypothesis, the study would have to compare the outcomes of those children who did not have an active father, and those who did. If there are different outcomes in children's well-being, some other way of understanding and describing the circumstances that might make that difference must be sought. By the same token, the distinction between semantic and affective distinctions cannot be posited as explanatorily useful in explaining different outcomes (e.g. different behavioural influences) if some so-called semantic associations investigated are also strongly affect-laden. In order to test the hypothesis, the study would have to compare associations which are not affect-laden with those which are. Until the distinction is deployed in a way that really does investigate distinct instances of each sort of implicit association, the findings can support the claim that the heterogeneity consists in two independent and distinct processes: affective and semantic.

Given these concerns, we do not here endorse the idea that what distinguishes different implicit associations (and any different underlying mechanisms) is that some are semantic and others affective. We should not endorse this as accurately capturing heterogeneous underlying processes involved in implicit biases. This is consequential for the notions that are at work in empirical psychology: we suggest that this distinction has been unsatisfactorily deployed. If Nosek is right that more fine-grained understandings of the cognitive processes involved in implicit cognition are needed, then so is more attention to the way that these processes are conceptualized and deployed in empirical studies.

Is it worth attempting to construct further studies which investigate this distinction? We have indications that different implicit associations generate different behavioural predictions. Whether this is a function of the affective content of the association, or indeed a distinct process particular to affective-laden associations, remains an open question—one worth pursuing insofar as it is

worth finding out what sorts of generalization can be made, and on what basis (the content of associations, or processes underlying them). It is not impossible to imagine how that distinction could be adequately operationalized. For instance, one might imagine an experimental paradigm in which the positive and negative words are replaced with (neuro)physiologically induced feelings of pleasure vs. displeasure on which to make more 'purely' affective categorizations. These could be contrasted with associations without affect (or with only a very low affective content). Such an undertaking is fraught with practical difficulties, but is not, in principle, impossible to implement. It is beyond the scope of this chapter to develop any full proposals for such efforts, but we would encourage psychologists and others to consider the possible ways of proceeding. Perhaps only if systematic content differences were then discerned might it be appropriate to consider claims about the underlying processes for the distinct content differences.

For present purposes, we need not establish that there are differences in terms of distinct and heterogeneous implicit processes. Rather, our aim is to draw attention to the fact that there are important functional differences with respect to some implicit associations—some of which are in terms of the degree of affect, which seems to make a difference (perhaps it is not the only thing that makes a difference) to the behavioural predictions generated. At present, experimental evidence supports the claim that implicit associations differ in some respects, such that some generalizations about implicit bias are unsupported. But, we have argued, it is not warranted to identify the respects in which they differ to be with regard to the associations being carved into two kinds: semantic or affective.

There remain two possibilities. One is to hold that implicit biases operate on fundamentally the same sort of process, but that they are dissociable and can functionally differ significantly in various dimensions (with respect to degree of awareness, relationship to explicit belief, behavioural predictions). Another is to hold that there are multiple processes involved in the category of implicit biases, and that these different processes correspond to the different features we have highlighted. At this stage, we do not believe that the considerations we have marshalled speak in favour of one or other of these theses—but there is much further work to be done on this topic.

3.3 Summary

We began by showing that experimental results illustrate functional differences which mean that it will be difficult to make certain generalizations about the phenomena that fall under the rubric of 'implicit bias'. These functional differences may be explicable in terms of content, or in terms of heterogeneity of underlying structure, such as whether the associations are affect-laden or not.

However, the way experimental studies have deployed the distinction between different kinds of association mean that those conclusions cannot provide support for there being two distinct kinds of process. It remains an open question what best explains the functional differences observed. That there are functional differences, however, is not in doubt. This claim is supported by the following considerations: the differences in content and the failure of correlation across implicit measures; the different relationship of implicit associations to explicit beliefs; and the different behavioural predictions generated by different implicit associations. To the extent that there is reason to believe that these functional differences are explained by heterogeneous cognitive processes, however, we suggest that further work is to be done in making precise the nature of this heterogeneity and deploying it in experimental design.

We have already alluded to the fact that these findings will have implications for philosophical discourse about implicit bias. In Section 4 we explain in more detail what we take these implications to be, and make specific recommendations about how philosophical discussions about implicit bias can accommodate these concerns.

4 Implications of the Heterogeneity of Implicit Bias

In this section we draw out the key implications of recognizing the functional heterogeneity of implicit bias.

4.1 Avoiding misleading generalizations, specific normative recommendations

The first implication of the aforementioned discussion pertains to the kinds of theoretical claims that have been made about implicit bias. Philosophers have reported on implicit bias in rather general terms, frequently talking of 'implicit bias' simpliciter or 'implicit race bias', rather than noting the particular kinds of association at issue. For clarity's sake, it would be useful to articulate the specific associations at issue. What particular stereotypical constructs are they associated with? Are evaluative associations at issue? What degree of negative affect is involved? One reason for which it is important to do so is that there are implications for the kinds of normative recommendations philosophers make about strategies for combating implicit bias.

Such strategies generally fall into two categories (see Jolls and Sunstein, 2006). Insulating strategies aim to put in place mechanisms that prevent bias from being activated by insulating individuals from the information that might activate them. For example, anonymizing CVs or essays means that evaluators do not have the information (about the gender, age, or race, and so on, of the evaluated

individual) that might trigger implicit associations that distort judgement or influence behaviour. This sort of strategy, therefore, need not be sensitive to the functional heterogeneity of implicit bias, insofar as it simply prevents bias triggering information from reaching the individual.

Mitigating strategies are those that attempt to limit any effects of implicit bias where activation and influence remains a possibility (because insulation from bias relevant information is not ~~possible~~). Mitigation might occur either by hindering the activation of the bias, or if it is activated, by blocking its influence upon judgement or action. For example, in interview contexts, where at least some salient social identities of an individual are not possible to 'cover up' or 'anonymize', steps need to be taken to reduce the likelihood of any implicit associations being activated, or if activated, from having an effect on judgement or action. Such steps might involve the deliberate exposure to counterstereotypical exemplars (so as to inhibit the activation of stereotypical associations) (Joy-Gaba and Nosek, 2010), or having agreed upon the weightings of criteria for evaluation (so that the influence of bias—which can lead merit to be redefined to accommodate bias—might be corrected) (Uhlmann and Cohen, 2005), or even pre-interview 'retraining' of behavioural disposition, so that avoidance dispositions are replaced with approach responses (Kawakami et al., 2007).

We argue that recommendations about the kind of mitigating strategies that should be undertaken need to be sensitive to the content of implicit associations likely to be at work, and the kind of behavioural outcome at issue. For example, the limited experimental findings outlined in Section 3.2 suggest that some implicit associations will influence judgement rather than approach/avoidance behaviours, and others will have greater influence on such behaviours (but less so on evaluative judgements). If this is right, then it is possible that a mitigating strategy might misfire by targeting an implicit association that is less likely to be influential in that particular context. For example, in light of the findings described in Section 3.2, we might say that if one is aiming to mitigate the influence of implicit associations on interracial interactions (which may involve approach/avoidance behaviours, such as seating distance) it would be a mistake to focus mitigating strategies on the implicit associations between mental/physical constructs and race (for example, by utilizing counterstereotypical exemplars to that stereotype). The strength of those associations did not correlate with greater seating distance (avoidance behaviour). Conversely, strategies which require individuals to reprogramme certain approach or avoidance responses to overcome implicit race biases might target negative affect (as has been shown in Kawakami et al., 2007) and make for smoother interracial interactions, but it is not clear that they will be effective in mitigating the influence of implicit

associations that feed into evaluations of competence or academic aptitude. For example, the IATs that measured the effects of approach/avoidance response reprogramming, in Kawakami et al. (2007), tested for impact on associations with generally positive (love, cheer, happy) and negative (pain, hate, evil) constructs, rather than for specific stereotypical associations such as those in Amodio and Devine (2006). Indeed, there is reason to suppose that certain strategies that may challenge negative affect would not be at all effective in mitigating stereotypical associations: the use of positive exemplars to challenge negative affect might nonetheless encode associations that affirm some other stereotypes. (For example, using Wilt Chamberlain or Michael Jordan as a positive exemplar might entrench the stereotypical associations concerning between black and physical, rather than mental, constructs).

Given the need for more information about the ways in which different implicit associations might operate differently, we are hesitant to make concrete proposals about how best to mitigate biases. Indeed, as more research reveals the different cognitive processes that may explain such functional differences, more research will be needed on what strategies are relevant to different implicit associations. However, our key claim is that it is important to be alert to the possibility that different associations are in play, and that adopting one strategy for mitigating implicit bias (e.g. exposure to counterstereotypical exemplars) is likely to be at best partial, and may address only ~~part~~ of the possible associations that could lead to implicitly biased outcomes. An awareness of how different strategies are effective in combating different implicit associations should counsel in favour of more comprehensive strategies for mitigating implicit biases.

Moreover, there are implications for individuals reflecting on whether they need to undertake such strategies. Precisely because implicit associations have been found to operate independently of each other, simply because an individual has been found not to have one implicit association (e.g. one IAT result that does not show an implicit race bias) does not mean that they do not have another quite similar one. As Alex Madva aptly expresses it (in correspondence): 'Maybe a given doctor has good interpersonal interactions with black people but still doesn't give them appropriate drug prescriptions. Just because you lack one racial bias doesn't mean you're off the hook.' Likewise, undertaking one bias-mitigating strategy does not mean that others will not remain operative. Recognizing the complexity of implicit associations, how they are related, and their functional heterogeneity, has important implications for evaluating one's own susceptibility to, and strategies for mitigating, the influence of bias.

4.2 Four recommendations

Research programmes into the way in which the different kinds of association differ and interact are ongoing. On the basis of the argumentation in this chapter, we first make the recommendation to empirical psychologists that the distinction between affective and semantic associations be revisited, and how it is experimentally deployed reconsidered. Moreover, because of the functional heterogeneity of implicit biases, and because of the difficulty of understanding how they might work together when multiple biases are all in play, it is important that the effects of implicit bias and interventions to tackle it be based on evidence beyond that garnered from psychology laboratories. We need to base recommendations for real-world interventions on rigorous field-experimental work.

We have three more recommendations for philosophers continuing to work on the range of important issues raised by the empirical findings about implicit bias. First, we recommend caution with respect to generalizations that are made about implicit bias. Whilst some generalizations are true and useful, we have drawn on evidence that indicates that other such generalizations are at best misleading.

Second, with respect to the formulation and implementation of normative claims concerning how to mitigate the effects of implicit biases, we recommend approaches that acknowledge the functional differences between implicit biases, and different strategies that might be needed to combat each of them. Attention to the different associations that might be involved in a given context, and the specific strategies that might be needed to combat the different kinds of implicit association, is needed. (We might aver that employing as many strategies as possible is the best plan, but, whilst a reasonable inference, this is as yet empirically unsupported.)

Finally, when writing about implicit bias, whilst the shorthand and general term 'implicit bias' can be useful, it would often be of helpful (both for assessing the truth of the claims made, and the likely efficacy of normative recommendations drawn from the claims) if the particular kinds of association at issue are articulated. This will assist in the identification of the association at issue, the contexts in which that particular association is likely to be particularly problematic, and the kinds of mitigating strategy that are likely to be efficacious. Recognizing and accommodating the heterogeneity of implicit bias may be an important step in effectively combating its effects.

References

Amodio, D. M. (2008). "The social neuroscience of intergroup relations." *European Review of Social Psychology* 19: 1–54

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.

Amodio, D. M., and Mendoza, S. A. (2010). "Implicit intergroup bias: Cognitive, affective, and motivational underpinnings." In Gawronski B. and Payne, B. K. (eds.), *Handbook of Implicit Social Cognition*. New York, NY: Guilford Press: 353–74.

Banaji, M. and Hardin, C. (1996). "Automatic stereotyping." *Psychological Science* 7: 136–41.

De Houwer, J. (2006). "What are implicit measures and why are we using them." In Wiers, R. W. and Stacy, A. W. (eds.), *The Handbook of Implicit Cognition and Addiction*. Thousand Oaks, CA: Sage: 1–28.

Dovidio, J. F. and Gaertner, S. L. (2000). "Aversive racism and selection decisions: 1989 and 1999." *Psychological Science* 11: 319–23.

Devine, P. G. et al. (2002). "The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice." *Journal of Personality and Social Psychology* 82(5): 835–48.

Fazio, R. H. (2007). "Attitudes as object-evaluation associations of varying strength." *Social Cognition* 25: 603–37.

Fazio, R. H., Jackson, J., Dunton, B., and Williams, C. (1995). "Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline?" *Journal of Personality and Social Psychology* 69: 1013–27.

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

Gallistel, C. R. (2008). "Learning and representation." In Menzel R. (ed.), *Learning Theory and Behavior*; vol. 1 of *Learning and Memory: A Comprehensive Reference*. Oxford: Elsevier: 227–42.

Gallistel, C. R. and King, A. (2009). *Memory and the Computational Brain: Why Cognitive Science will Transform Neuroscience*. New York: Blackwell/Wiley.

Gallistel, C. R., Matzel, L. D. (2013). "The neuroscience of learning: Beyond the Hebbian synapse." *Annual Review of Psychology* 64: 169–200.

Garcia, J. L. A (1996). "The heart of racism." *Journal of Social Philosophy* 27(1): 5–46.

Gawronski, B., Hofmann, W., and Wilbur, C. J. (2006). "Are 'implicit' attitudes unconscious?" *Consciousness and Cognition* 15: 485–99.

Gawronski, B. and Payne, B. K. (eds.) (2010). *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York: Guilford Press.

Gendler, T. S. (2011). "On the epistemic costs of implicit bias." *Philosophical Studies* 156(1): 33–63.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., and Mellott, D. S. (2002). "A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept." *Psychological Review* 109: 3–25.

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

Haslanger, S. (2008). "Changing the ideology and culture of philosophy: Not by reason (alone)." *Hypatia* 23(2): 210–22.

Holroyd, J. (2012). "Responsibility for Bias." *Journal of Social Philosophy* Special issue, ed. Crouch, M. and Schwartzman, L.

Holroyd, J. (2014). "Implicit bias, awareness and epistemic innocence." *Consciousness and Cognition* Special Issue on Imperfect Cognitions, ed. Bortolotti, L. and Bissett-Sullivan, E.

Jolls, C. and Sunstein, C. (2006). "The law of implicit bias." *California Law Review* 94: 969–96.

Joy-Gaba, J. A. and Nosek, B. A. (2010). "The surprisingly limited malleability of implicit racial evaluations." *Social Psychology* 41: 137–46.

Kawakami, K., Phills, C. E., Steele, J. R., and Dovidio, J. F. (2007). "(Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors." *Journal of Personality and Social Psychology* 92(6): 957–71.

Kelly, D. and Roedder, E. (2008). "Racial cognition and the ethics of implicit bias." *Philosophy Compass* 3(3): 522–40.

Machery, E., Faucher, L., and Kelly, D. (2010). "On the alleged inadequacies of psychological explanations of racism." *The Monist* 93: 228–54.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., and Jetten, J. (1994). "Out of mind but back in sight: Stereotypes on the rebound." *Journal of Personality and Social Psychology* 67(5): 808–17.

Monteith, M. and Voils, C. (1998). "Proneness to prejudiced responses: Toward understanding the authenticity of self-reported discrepancies." *Journal of Personality and Social Psychology* 75(4): 901–16.

Monteith, M. J., Voils, C. I., and Ashburn-Nardo, L. (2001). "Taking a look underground: Detecting, interpreting and reacting to implicit racial biases." *Social Cognition* 19(4): 395–417.

Nosek, B. A., Hawkins, C. B., and Frazier, R. S. (2012). "Implicit social cognition." In Fiske, S. and Macrae, C. N. (eds.), *Handbook of Social Cognition*. New York, NY: Sage: 31–53.

Nosek, B., Greenwald, A., and Banaji, M. (2007). "The Implicit Association Test at age 7: A methodological and conceptual review." In Bargh, J. (ed.), *Automatic Processes in Social Thinking and Behaviour*. New York, NY: Psychology Press: 265–92.

Olson, M. A. and Fazio, R. H. (2004). "Reducing the influence of extrapersonal associations on the implicit association test: Personalizing the IAT." *Journal of Personality and Social Psychology* 86: 653–67.

Payne, B. K., Cooley, E., Loersch, C., and Lei, R. (ms.). "Who owns implicit attitudes? Testing a meta-cognitive perspective."

Saul, J. (2013) "Implicit bias, stereotype threat and women in philosophy." In Jenkins F. and Hutchison, K. (eds.). *Women in Philosophy: What Needs to Change?* Oxford: Oxford University Press: 39–60.

Uhlmann, E. L. and Cohen, G. L. (2005). "Constructed criteria redefining merit to justify discrimination." *Psychological Science* 16(6): 474–80.

Valian, V. (2005). Beyond Gender Schemas: Improving the Advancement of Women in Academia, *Hypatia* 20(3): 198–213.

Webb, T. L., Sheeran, P., and Pepper, J. (2010). "Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials." *British Journal of Social Psychology* 51(1): 13–32.

Wegener, D. T., and Petty, R. E. (1995). "Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias." *Journal of Personality and Social Psychology* 68: 36–51.