# What We Can (And Can't) Infer About Implicit Bias From Debiasing Experiments

Nick Byrd
Departments of Philosophy and Psychology, Florida State University, Tallahassee, FL
Google Scholar: https://scholar.google.com/citations?user=d3prs2YAAAAJ&hl=en
ORCID: 0000-0001-5475-5941
nick.a.byrd@gmail.com

**Abstract**

The received view of implicit bias holds that it is associative and unreflective. Recently, the received view has been challenged. Some argue that implicit bias is not predicated on "any" associative process, and it is unreflective. These arguments rely, in part, on debiasing experiments. They proceed as follows. If implicit bias is associative and unreflective, then certain experimental manipulations cannot change implicitly biased behavior. However, these manipulations can change such behavior. So, implicit bias is not associative and unreflective. This paper finds philosophical and empirical problems with that argument. When the problems are solved, the conclusion is only half right: implicit bias is not necessarily unreflective, but it seems to be associative. Further, the paper shows that even if legitimate non-associative interventions on implicit bias exist, then both the received view and its recent contender would be false. In their stead would be interactionism or minimalism about implicit bias.

**keywords**: debiasing, dual process theory, implicit bias, implicit association test, associationism, reflectivism, interventionism, philosophy of mind, philosophy of cognitive science, philosophy of science

**Acknowledgements**. Thanks to Mike Bishop, Mike Dacey, Bryce Huebner, Luis Rosa, John Schwenkler, anonymous reviewers, and editors for comments on earlier versions of this paper. Thanks to Istvan S. N. Berkeley, John Bickle, David Chalmers, Gabriel De Marco, Grace Helton, Zoe Jenkin, Eric Mandelbaum, Michele Merritt, Valentina Petrolina, Jake Quilty-Dunn, Susanna Siegel, and Evan Westra for shrewd comments on previous presentations of this paper. Thanks to Cameron Buckner, Bertram Gawronski, Angela Smith, and Ege Yumusak for helpful personal correspondence about this project more generally.

**Funding**. This research was supported, in part, by a graduate assistantship from the Graduate School at Florida State University, by travel funding from the Congress of Graduate Students at Florida State University, by travel funding from the Department of Philosophy at Florida State University, and by a Graduate Student Travel Award from the American Philosophical Association.

*The imagination is influenced by associations of ideas; which, …are not easily altered.*

David Hume (1983)

Imagine nutrition scientists discover that bodyweight can be changed not only by calorie ingestion and consumption, but by other factors. When science columnists catch wind of these findings, they write up pieces with titles like "Why Calories Don't Matter", arguing that gaining and losing weight is not predicated on "any" caloric processes. Some columnists go as far as to recommend that the received, thermodynamic view of bodyweight be abandoned. Obviously, the science columnists' conclusions do not follow. The scientists did not demonstrate that changes in bodyweight are not predicated on *any* caloric processes. Rather, the scientists demonstrated that some weight changes are not predicated on "only" caloric processes. That finding is consistent with the idea that bodyweight is predicated on caloric processes, even if not fully. This paper cautions against the science columnists' any-only mix-up when thinking about implicit bias: the mistake of concluding that implicit bias is not predicated on *any* instances of a particular process when the evidence merely shows that implicit bias is not predicated on *only* instances of that particular process.

Discussions of implicit bias are increasingly common. Debate moderators ask presidential candidates about implicit bias (Blake, 2016), Fortune 500 companies close thousands of stores in order to teach their employees about implicit bias (Meyer, 2018), and philosophers worry that implicit bias poses epistemic threats to philosophy (e.g., Saul, 2013a, 2013b; Peters, *forthcoming*). Nonetheless, some are skeptical about the existence of implicit bias or the efficacy of corporate implicit bias training (e.g., McCoy, 2018). So, academics try to remind the public about evidence of implicit bias (e.g., Payne, Niemi, & Doris, 2015) and successful debiasing (e.g., Carley, 2018). Philosophers of mind have taken this evidence seriously, arguing that these debiasing findings undermine the received view of implicit bias (e.g., Mandelbaum, 2016) and demand new solutions to implicit bias (e.g., Huebner, 2016; Madva, 2017; Saul, 2013a).

Given these stakes in philosophy and in public discourse, one will want to take every opportunity to be careful about what they infer about implicit bias from debiasing experiments. This paper explains how to identify methodologically sound debiasing experiments and determine what they tell us about implicit bias. Section 1 explains and distinguishes nine views of implicit bias. Section 2 explains how to (and how not to) draw inferences from debiasing experiments. Then, Section 3 reviews influential debiasing experiments, highlighting differences in methodological quality along the way. Section 4 explains what follows from the strongest evidence, using the inference principles from earlier sections. Of course, a paper this size cannot carefully examine every debiasing experiment. So, Section 4 also explains what would follow if forthcoming or overlooked debiasing experiments' findings differ from the findings considered herein. The primary conclusion is that up to three views of implicit bias are compatible with current and future evidence: associationism, interactionism, or minimalism. A secondary conclusion is a sort of reflectivism about implicit bias. These conclusions imply that both the received view and more recent non-associationist views of implicit bias are incompatible with strong evidence. Reviewing some of the literature on implicit bias will help explain how these conclusions follow.

## 1 IMPLICITLY BIASED BEHAVIOR

The most well-known measure of implicitly biased behavior is the Implicit Association Test (IAT for short). The IAT is a categorization task. Various versions of the test measure various modes of implicit biases in behavior. For example, the Race IAT measures differences in responses to racial stimuli. This paper will focus on the Race IAT, but its analysis can be fruitfully applied to other versions of the IAT and other indirect measures of bias (see Appendix).

The IAT includes multiple phases of categorization. In the first phase of the Race IAT, participants press buttons on a keyboard to categorize words into one of two categories: GOOD or BAD. Then

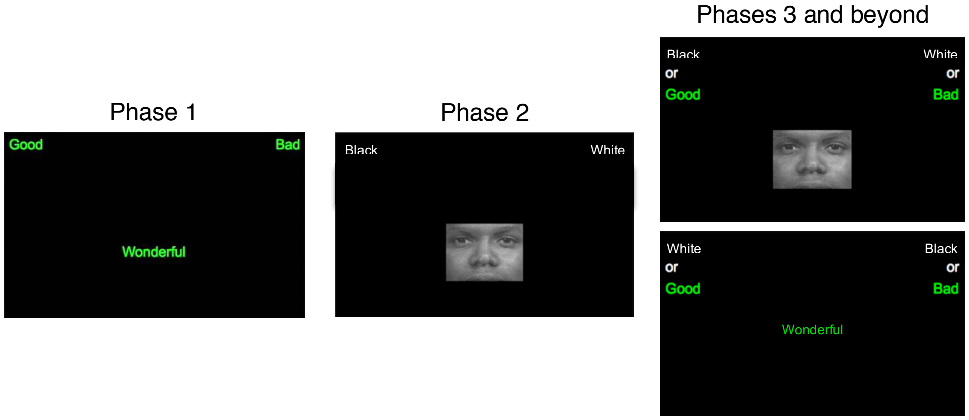

*Figure 1.* Phases of the Implicit Association Test: word categorization, face categorization, and word-and-face categorization.

participants categorize faces with either white or black racial features into one of two categories: WHITE or BLACK (Figure 1). This much is fairly straightforward.

In each subsequent phase, participants categorize *either* faces *or* words, one at a time, into composite categories: In one phase, the composite categories might be BLACK/GOOD or WHITE/BAD and in the following phase, composite categories might be WHITE/GOOD or BLACK/BAD. It is in these latter phases with composite categories where interesting patterns emerge. Most participants' categorization accuracy and response latencies reveal a preference for white facial features over black facial features (e.g., Greenwald, McGhee, & Schwartz, 1998). That is, participants are quicker to pair black facial features than white racial features with composite categories containing BAD. And, likewise, participants are quicker to pair white facial features with composite categories containing GOOD.

It is not uncommon to detect such implicit Pro-White biases in the behavior of those who explicitly express Pro-Black preferences (e.g., Gaertner & McLaughlin, 1983). While this does not suggest that people are unaware of their own biases (Gawronski, Hofmann, & Wilbur, 2006; Gawronski, *forthcoming*), it does suggest that behavior can be biased in ways that are not consciously endorsed or even in ways that are consciously disavowed.

Naturally, this disconnect between implicit biases in behavior and more explicit attitudes might raise questions about whether there is a disconnect between implicit biases and behaviors besides button-pressing (Greenwald, Andrew, Uhlmann, & Banaji, 2009; Greenwald, Banaji, & Nosek, 2015; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2015). In short, one might wonder about the validity of measures like the IAT. The virtue of the IAT is its ability to accurately quantify error rates and reaction times and other indirect measures of attitudes and behavior in controlled settings (Jost, 2018). More ethologically valid measures of implicit biases in behavior make quantification, timing, and control more challenging—e.g., implicit biases in resume evaluation (e.g., Tyler & Mccullough, 2009) and seating distance (Sechrist & Stangor, 2001). Fortunately, the present paper's analysis will apply to debiasing according to any indirect measure of biases in behavior. So, concerns about the validity of the IAT undermine the present investigation only if these concerns generalize to all indirect measures.

The name 'Implicit Association Test' advertises how implicitly biased behavior was initially thought to be predicated on associations (Greenwald et al., 1998). Consequently, this associative view of implicit bias became the received view of implicit bias among philosophers (e.g., Gendler, 2008a, 642; 2008b, 577). Philosophers describe associations as "pairs of thoughts [that] become associated based on […] past experience" (Mandelbaum, 2017). Accordingly, the associative explanation of the Race

IAT findings is roughly as follows: people experience White racial features paired with positive valences more than negative valences and they experience Black racial features paired with negative valences more often than positive valences. This conditioning results in associations between White racial features and positive valences or Black racial features and negative valences. These associations explain why unendorsed preferences for certain racial stimuli would manifest on tasks like the IAT.

However, this associative view of implicit bias has become controversial. Some argue that implicit bias is belief-like (Mandelbaum, 2013; cf. Madva, 2015) and that implicit bias is "not predicated on *any* associative structures or processes" (Mandelbaum, 2016, p. 629). Others argue that while implicit bias might be belief-like, such beliefs are nonetheless dispositional (Schwitzgebel, 2002, 2010; cf. Quilty-Dunn & Mandelbaum, 2017). Yet others argue that implicit bias is less like belief and more like a patchy endorsement (Levy, 2015) or a trait (Machery, 2016). And, coming full circle, some admit that implicit bias might be associative after all, even if only in part (e.g., De Houwer, 2006; Del Pinal & Spaulding, 2018, Huebner, 2016; Gawronski & Bodenhausen, 2014). Some background theory and evidence will explain why anyone would want to abandon the received, associative view of implicit bias for other views.

## 1.1 Dual Process Theory

Consider the dual-process theory of cognition. The theory distinguishes between at least two types of processes with labels such as 'Type 1' and 'Type 2' (e.g., Evans & Stanovich, 2013; Table 1) or 'System 1' and 'System 2' (e.g., Evans, 2009, Table 2.1; Frankish, 2010, Table 1). To make it easier to remember what these labels describe, this paper will borrow more informative labels for each type of processing: Type 1 processes will be labeled 'non-reflective' and Type 2 processes 'reflective' (*à la* Pennycook, Cheyne, Koehler, & Fugelsang, 2015; Strack & Deutsch, 2004). Some common dual-process distinctions are found in Table 1.

*Table 1.* Dual Process Descriptions

| Non-reflective (Type 1) | Reflective (Type 2) |
|---|---|
| associative | non-associative |
| fast | slow |
| automatically processed | deliberately processed |
| not consciously represented | consciously represented |

Of course, one need not buy all the common dual process distinctions—at least, not without qualification. Indeed, one might be suspicious of binary distinctions in psychology more generally (Newell, 1973). Fortunately, one need not accept all common or binary dual-process distinctions in order to accept the conclusions of this paper. Consider two examples of common dual-process theory distinctions that need not be accepted without qualification.

Start with the associative vs. non-associative distinction. Explaining behavior in terms of associations is about as old as philosophy (Anderson & Bowen, 1980, 9), so many construals of associations have accumulated. Hume thought that associations operate automatically and unconsciously.

> Tis evident, that the association of ideas operates in so silent and imperceptible a manner, that we are scarce sensible of it, and discover it more by its effects than by any immediate feeling or perception (Hume, 1978).

Some cognitive scientists have adopted such Humean construals of associations. For example;

> When a response is produced solely by the associative system, a person is conscious only of the result of the computation, not the process. Consider an anagram such as 'involnutray' for which the correct answer likely pops to mind associatively (involuntary) (Sloman, 1996, 6).

However, the Humean construal of associations is controversial. Indeed, there are plenty of reasons to think that associations can cross the conscious/non-conscious divide (Dacey, 2016; Devine, 1989; Fridland, *forthcoming*; Hahn, Judd, Hirsch, & Blair, 2014). Because of this, some have cautioned against inferring either that cognitive processing is necessarily associative because it is automatic or unconscious or that it is necessarily automatic and unconscious because it is associative (Mandelbaum, 2016, p. 647; cf. Hütter & Sweldens, 2018). Importantly, this implies that the associative vs. non-associative distinction could be orthogonal to the non-reflective vs. reflective distinction (*contra*, for example, Strack & Deutsch, 2004). This paper takes that possibility seriously, as I explain below.

Consider the distinction between fast and slow processing (e.g., Kahneman 2011), which is also controversial. Seemingly reflective reasoning is sometimes fast (Bago & De Neys, 2017). For this and other reasons, many cognitive scientists seem to reject a definite distinction between fast and slow processing (Krajbich, Bartling, Hare, & Fehr, 2015; Pennycook, Fugelsang, Koehler, & Thompson, 2016; Sun, 2016). However, one can admit that the boundary between fast and slow is vague while maintaining that there is a range of response times within which mental representations are unlikely to be available for conscious control or even explicit endorsement (Posner & Snyder, 1975).

At this point, a critic of dual-process theory might begin to question the existence or utility of a dual-process distinction (Melnikoff & Bargh, 2018). However, the critic should remember that the absence of a clear categorical dual-process distinction does not show that dual-process distinctions are altogether illegitimate (Pennycook, Neys, Evans, Stanovich, & Thompson, 2018). A categorical distinction proposes a clear boundary between two concepts, whereas a comparative distinction merely proposes a relative difference between two concepts (Carnap 1950, Section 3 to 8). So, dual-process theorists have explicated some dual-process distinctions comparatively rather than categorically (e.g., Evans & Stanovich, 2013, 229-231). That brings us to the two dual-process distinctions employed in this paper.

First, this paper will employ the common distinction between reflective and non-reflective processing. However, this distinction will be comparative rather than categorical. Reflective processing is more consciously represented and deliberately processed while non-reflective processing is less consciously represented and more automatically processed (Shea & Frith, 2016). Cognition is more conscious when participants are more aware of, more able to articulate, and/or more able to process it at the personal level (ibid.). Cognition is more deliberate when it involves more interruption of or less acceptance of the output of automatic processing (Bargh, 1992; Fridland, 2016; Moors & De Houwer, 2006). This explication of reflection will be familiar to anyone who is aware of the famous cases of reflection from philosophy and psychology: someone finds their first intuition plausible, but steps back for a moment to consider their intuition, and then either endorses the intuition or arrives at a new response (e.g., Frederick, 2005; Korsgaard, 1996).

Second, I will employ a categorical distinction between associative and non-associative processing. Before I describe this categorical distinction, two caveats are in order. First, while processing is either associative or non-associative, attitudes and behavior may not be so binary. Indeed, one of the morals of this paper will be that one and the same behavior can be influenced by both associative and non-associative processes. Second, there is an emerging literature which disputes what associative processing can and cannot do (e.g., Buckner, 2017; cf. De Houwer, 2018). Since that debate has yet to resolve, I will grant a conventional notion of associative processing and point interested readers toward the unfolding debate (Corneille & Stahl, 2018). Conventionally, cognitive processing is associative if it can be well-described by stimulus-response phenomena such as conditioning or counterconditioning (*à la* Mandelbaum, 2016). Conditioning and counterconditioning involve repeatedly activating two representations until activating one representation also activates the other representation. This explication of associations captures the kind of processing that might be involved in the behavior that is measured by the Race IAT. For example, a racial association might be formed as follows. For whatever reason, someone repeatedly experiences BLACK MALE paired with DANGER. These experiences create and strengthen an association between the concept representation (BLACK MALE) and the negatively valenced representation (DANGER). Once the association is formed, the mere activation of BLACK MALE activates the negative valence DANGER. That automatic activation of negative valence is supposed to explain the often-unendorsed reflexive biases that manifest during the Race IAT.

A 2x2 matrix can be constructed to sort cognition according to the two distinctions just explained (Figure 2). The boundary between the left and right sides of the matrix separates associative from non-associative processing. The fuzzy boundary between the top and bottom separates more reflective from less reflective processing.

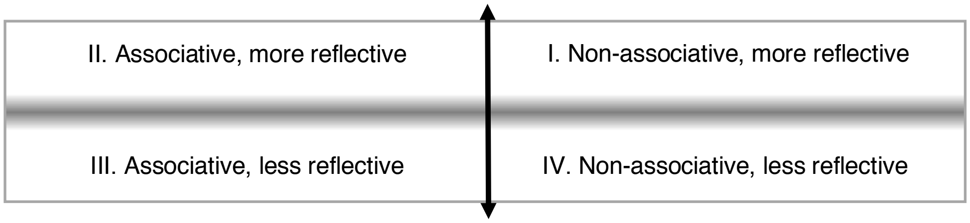
II. Associative, more reflective
I. Non-associative, more reflective
III. Associative, less reflective
IV. Non-associative, less reflective

*Figure 2*: Matrix distinguishing four modes of cognition.

One might think that this deviates from dual-process theory since it proposes four processes. In reality, this merely proposes that two common dual-process distinctions are orthogonal. Besides, this more-than-two quadrant approach to dual-process theory is already common among cognitive scientists (e.g., Evans, 2009; Stanovich, 2009; Gawronski & Bodenhausen, 2014; Shea & Frith, 2016). Further consideration of these approaches and dual-process theory goes beyond the scope of the present investigation. The existing conceptual space need only represent key differences between the views of implicit bias under consideration in the present paper.

## 1.2 Nine Views of Implicit Bias

The matrix just described allows us to classify views of implicit bias based on whether they predicate implicit bias on (1) associative or non-associative processing as well as (2) more or less reflective processing. This produces a 3x3 matrix of nine categories of views about implicit bias (Table 2).

*Table 2*. A matrix of up to nine modes of cognitive processing on which implicit bias could be predicated.

| **Implicit bias is predicated on…** | **Associative processing** | **Associative or non-associative processing** | **Non-associative processing** |
|---|---|---|---|
| **More reflective processing** | Reflective associationism about implicit bias | Reflective interactionism about implicit bias | Reflective non-associationism about implicit bias |
| **More or less reflective processing** | Associationism about implicit bias | Interactionism about implicit bias | Non-associationism about implicit bias |
| **Less reflective processing** | The received view of implicit bias | Unreflective interactionism about implicit bias | Unreflective, non-associationism about implicit bias |

In the left column are associationist views of implicit bias. In the bottom, left cell is the received view of implicit bias, claiming that implicit bias is predicated on associations that are processed less reflectively. The cell above the received view of implicit bias refers to a more capacious associationism about implicit bias, according to which implicit bias is predicated on associations, but allowing that associations can be processed more reflectively or less reflectively.

In the middle column are interactionist views about implicit bias. Interactionism about implicit bias claims that implicit bias can be predicated on associated and non-associative processes. Interactionists about implicit bias can, in principle, claim that implicit bias is predicated on more reflective processing (top center), less reflective processing (bottom center), or some combination thereof (middle, center).

In the far-right column are non-associationist views about implicit bias. Non-associationism about implicit bias can be described by statements like, "implicit biases are not predicated on any associative structures or associative processes" or "the structure of implicit bias is not, after all, underwritten by associations" (Mandelbaum 2016, pp. 629, 637). Similar to associationism and interactionism about implicit bias, non-associationism about implicit bias can vary depending on whether it predicates implicitly biased behavior on more reflective processing, less reflective processing, or some combination thereof.

Before we determine how to infer these views of implicit bias, a few words of clarification are in order. First, determining who has defended each kind of view is a worthy historical project, but that is beyond the scope of the present investigation. Second, this matrix might not classify all possible views of implicit bias. Third, the matrix spares some of the details of the views that it classifies. For example, the matrix's interactionist views specify *that* associative and non-associative processes interact, but do not specify how these processes interact—the latter is discussed in the next section and by Gawronski & Bodenhausen (2014). Nonetheless, the matrix visualizes various answers to an ongoing question in the debate about the nature of implicit bias: Should the received view of implicit bias be abandoned for a more centrist, more far-right, or more reflective view of implicit bias?

## 2 INFERENCES FROM DEBIASING EXPERIMENTS

Determining the kind(s) of processing on which implicit bias is (and is not) predicated involves determining the kinds of processing on which debiasing is (and is not) predicated. In other words, views of implicit bias depend—at least in part—on what can be inferred from debiasing experiments. So, we need to determine what can be inferred.

### 2.1 Existing Inferences from Manipulation

One way to proceed is to follow precedent. The debiasing literature contains at least two inferential principles for determining the types of processing on which implicit bias is (and is not) predicated. One common inference in the debiasing literature assumes the following principle of affirmation.

> *Affirmative Manipulation Principle*. S is predicated on P-type processing just in case a P-type manipulation changes S.

The affirmative manipulation principle seems to feature in hypothetical deductions that implicit bias is propositional.

> …if you find two negatives making a positive, what you've found is a propositional, and not an associative, process. […] When a person you don't like dislikes another, you tend to like that other person. When a person you don't like dislikes another, you tend to like that other person. So, a negative valence when combined with a negative valence somehow results in a positive valence. The 'somehow' […] is sensible on a propositional theory. (Mandelbaum 2016, pp. 640-641)

The point is not that two positives making a negative cannot be well-explained by associative processes. Like many claims about what counts as an associative process (e.g., De Houwer, 2018), that point is controversial (Toribio 2018b; see also Cone, Mann, & Ferguson, 2017;). For example, Gawronski, Walther, and Blank (2005) do not endorse that claim when reporting that multiple positives made a negative. Also, that claim relies on a dichotomy between associative and propositional processing that might be false given what is already known about human and animal psychology (Buckner 2017, 3-6). Hence, the point is just that if we agree that an experimental manipulation involves certain types of processing, then we should agree that any phenomena changed by that manipulation are predicated on those types of processing. I will grant the legitimacy of the affirmative inference principle in this paper.

Another common inference in the debiasing literature assumes the following principle of negation.

> *Negative Manipulation Principle*. S is *not* predicated on P-type processing just in case S is manipulated by a non-P-type manipulation or S is not manipulated by a P-type manipulation.

The negative manipulation principle seems to be at work in arguments for non-associationism about implicit bias. For example;

> … if AIB [associationism about implicit bias] is true, then no logical or evidential interventions should directly work to change implicit attitudes. […] If there are [such] interventions that reliably work to counteract implicit bias […], then we have evidence that the structure of implicit bias is not, after all, underwritten by associations. […. And] a logical intervention did in fact have an impact on participants' implicit attitudes. (Mandelbaum 2016, pp. 635, 637, 645)

The conclusion is that "we have evidence that the structure of implicit bias is not, after all, underwritten by associations" (ibid.) The idea is that logical or evidential manipulations are non-associative. So, according to the negative manipulation principle, if non-associative manipulations change implicit biases, then those implicit biases are not predicated on associations.

However, there are problems with the negative manipulation principle. One problem is that the negative manipulation principle leads to something like the science columnists' any-only mix-up: the mistake of thinking that implicitly biased behavior cannot be predicated on any associative process because implicitly biased behavior is not predicated on only associative processes.

## 2.2 Manipulation Is Not Enough for Negative Inference

To avoid the science columnists' any-only mix-up about implicit bias, the negative manipulation principle will need to be replaced with a more circumspect principle, like the one below.

> *Negative Intervention Principle*. S is not predicated on P-type processing just in case both P-type manipulations or measurements and non-P-type manipulations or measurements are employed and, empirically, only non-P-type cause a change in S.

Unsurprisingly, the difference between the negative manipulation principle and the negative intervention principle has to do with the difference between manipulation and intervention. Not all philosophers or social scientists distinguish manipulation from intervention—indeed, many use the words interchangeably. So, a definition of 'intervention' is in order: process P intervenes on S only when the change in S is caused by only P. Or, in causal graph terminology, P intervenes on S only when it "breaks all other arrows directed into" S (Woodward, 2016).

Manipulations, on the other hand, change S in a way that can involve multiple causes. So, manipulations do not show that a change was *caused by only one process*. However, interventions show both: that something is changed and that the change was caused by only one process. So, while manipulations are necessary for intervention, they are not sufficient for intervention because interventions are a subset of manipulations. By conjunction, we are less likely to detect interventions than manipulations. This can be illustrated by imagining the development of a debiasing research program.

**Exploratory experiments**. At first, researchers just want to see if any manipulation whatsoever can cause debiasing. So, researchers do not design their experimental manipulations according to the theoretical likelihood that they involve a particular type of processing such as associative or non-associative processing. Rather, researchers design their manipulations based on anecdotes or intuitions about how debiasing works. Eventually, the researchers find that certain experimental conditions reduce implicit biases in behavior significantly more than control conditions.

Because the researchers do not have strong theoretical reasons to think that their manipulation involved only associative or only non-associative processing, there are at least six viable interpretations of their debiasing results (Figure 3, adapted from Figure 1 in Perugini, Richetin, & Zogmaister, 2010; see also Madva, 2015, Section 6, and Brownstein, 2018 for more discussion of this interpretive difficulty). Only two of these six interpretations involve a single type of processing intervening on implicitly biased behavior (3a and 3b). The other four interpretations involve two kinds of processing jointly manipulating implicitly biased behavior (3c, 3d, 3e, and 3f). So, exploratory experiments cannot arbitrate between the various views of implicit bias from Table 2.

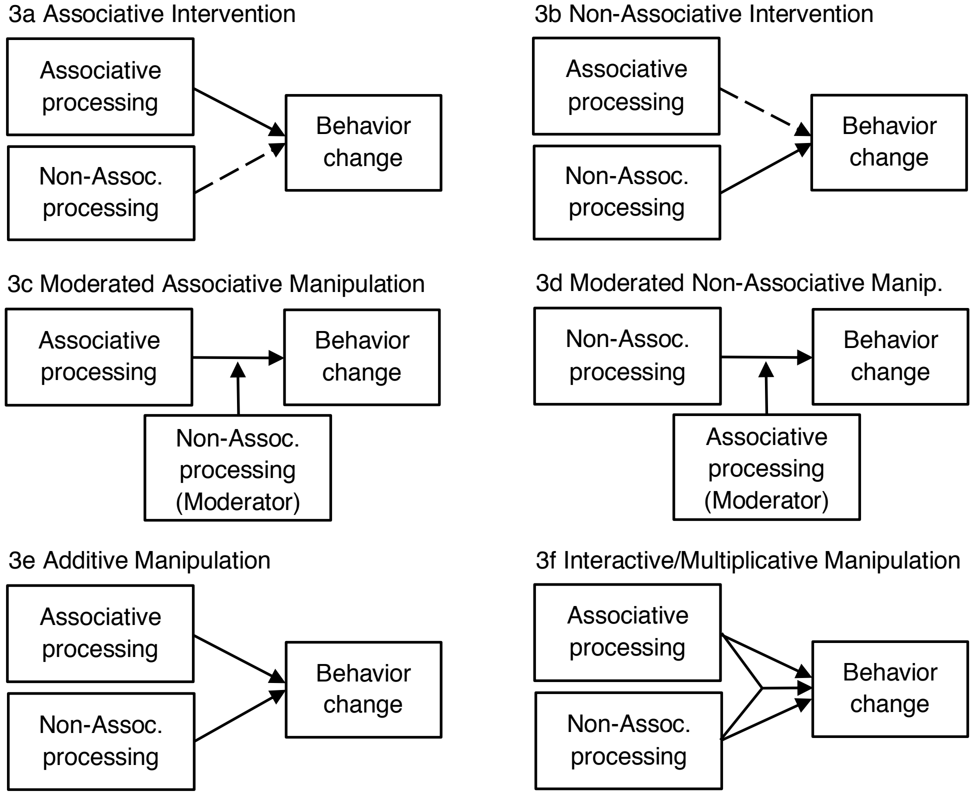

*Figure 3*. Intervention patterns (3a and 3b) vs. manipulation patterns (3c, 3d, 3e, and 3f) adapted from Perugini and colleagues (2010, Figure 1). NOTE: dashed lines denotes a broken causal arrow.

**Univariate follow-up experiments**. After the exploratory experiments, researchers decide to manipulate something that is widely accepted to involve only one type of processing. They try to manipulate implicitly biased behavior via associative processing by presenting participants with counterstereotypes (i.e., counterconditioning). Analyses of their data reveals that their associative manipulations repeatedly reduced implicit biases in behavior compared to their control groups.

A few things follow from these results. First, these univariate findings negate only one of the six interpretations from Figure 3 (i.e., 3b). Second, via the affirmative manipulation principle, researchers can infer that implicitly biased behavior is predicated on associative processing. However, they cannot infer that implicitly biased behavior is not predicated on non-associative processing. Indeed, the researchers did not measure (or manipulate) non-associative processing. So, they cannot analyze the impact of non-associative processing. Therefore, while univariate follow-up experiments tell us something about the nature of implicit bias, they hardly settle the debate about the nature of implicit bias.

**Multivariate follow-up experiments**. After our univariate experiments, researchers decide to detect interventions on implicitly biased behavior via either associative or non-associative processing. So, they design two kinds of manipulations. The associative manipulation attempts to change implicit biases in behavior by presenting participants with counterstereotypes (i.e., counterconditioning). The non-associative manipulations attempt to change implicit biases in behavior in a way that is compatible only with non-associative processing—see Mandelbaum (2016) and De Houwer (2018) for discussions of such manipulations. Then the researchers randomly assign participants to associative manipulation conditions, non-associative manipulation conditions, or a control group.

Now, imagine what we can infer if all multivariate follow-up experiments find that only one kind of manipulation condition reduces implicitly biased behavior significantly more than the control condition. First, we can negate all but one of the interpretations from Figure 3 (i.e., either 3a or 3b would remain). Second, we can infer—via the negative intervention principle—that implicit biased behavior is not predicated on *any* of one type of processing—i.e., that it is predicated entirely on another type of processing. Thus, multivariate follow-up experiments have the potential to settle the debate about the nature of implicit bias.

However, imagine what we can infer if both manipulation conditions reduce implicitly biased behavior significantly more than the control condition. First, we can negate only two of the six interpretations in Figure 3 (i.e., 3a and 3b). Second, via the affirmative manipulation principle, we can infer that implicitly biased behavior can be predicated on both associative and non-associative processing. In other words, while univariate follow-up experiments allow us to infer something about the nature of implicit bias, they will not necessarily settle the existing debate about the nature of implicit bias.

## 2.3 What to Infer from Null Results

Another possibility is that debiasing experiments find no manipulations that result in long-lasting, reliable, and significant reductions in implicit bias compared to controls. This is not a far-fetched possibility. Some meta-analyses find that experimental manipulations of implicit bias are "relatively weak" (Forscher et al., 2018).

Weak or even null results in debiasing experiments are to be expected if the average correlation between two IAT scores from the same person, i.e., test-retest reliability, is not high (e.g., $0.45 < r < 0.63$ in Bar-Anon & Nosek, 2014; average $r = 0.54$ in Gawronski, Morrison, Phill, & Galdi, 2017). One might think that this would show that the IAT is an unreliable measure. However, IAT scores could have high internal consistency (e.g., 0.83 and .88, ibid.) even if test-retest reliability is low. So, one interpretation of these findings could be that the IAT reliably measures something that is not highly stable over time (Gawronski, *forthcoming*; Jost, 2018). Insofar as that is right, it will be more difficult to detect changes in implicit bias that are the result of debiasing manipulations rather than the result of ordinary instability in implicit bias. Further, insofar as that is right, it will be more difficult to infer anything about the nature of implicit bias from debiasing experiments alone.

## 3 DEBIASING EXPERIMENTS

Philosophers sometimes cite debiasing experiments in favor of their view about implicit bias. As I will argue below, the evidential value of these experiments varies. A few experiments are under-described, making their evidential value indeterminable. Other experiments are adequately described but do not address several methodological concerns, mitigating their evidential value. Only some experiments are scrupulous enough to constitute strong evidence. Naturally, one should infer views of implicit bias from the strong evidence.

### 3.1 Under-described Evidence

Mandelbaum mentions a debiasing experiment which found that variations in argument strength can manipulate implicitly biased behavior (2016, p. 640). The experimenters presented an unspecified quantity of undergraduates with either strong or weak reasons in favor of a new policy to integrate more black professors at their university (Briñol, Petty, & McCaslin, 2009, p. 293). The strong reasons were as follows: "the number and quality of professors would increase with this program (without any tuition increase) [and] the number of students per class would be reduced by 25%" (ibid., 294). The weak reasons were as follows: "the program would allow the university to take part in a national trend and with the new professors, current professors might have more free time to themselves" (ibid.). After participants were presented with these strong or weak reasons for the policy, they were given the Race IAT.

Briñol and colleagues found that participants who were presented with strong reasons for the pro-Black policy were more positive toward Black facial features than participants who received weak reasons. The authors consider this an associative manipulation. However, some argue that this finding would be difficult to explain via only associative processing (e.g., Mandelbaum, 2016). If that is right, then Briñol and colleagues' manipulation would be non-associative. Alas, even if we grant that, we do not yet have enough information about the finding to know if we should infer anything from it. While this experiment's design has promise, its sample size and other descriptive statistics (e.g., the p-value and the effect size) are not reported—cf. Horcajo, Briñol, & Petty, 2010 in Appendix for similar experiments with descriptive statistics about non-racial stimuli. Historically, not reporting such relevant details has been common in some social sciences (McCloskey & Ziliak, 1996). However, unless or until the details of such under-described experiments are provided, their evidential significance is indeterminable (ibid.).

### 3.2 Mitigated Evidence

Mandelbaum also references a debiasing experiment which found that differences in peer disagreement can manipulate implicitly biased behavior (2016, p. 641). The experimenters sorted about 50 undergraduate psychology students into a low-bias group and a high-bias group based on their level of racial bias (Sechrist & Stangor, 2001). Once sorted, each group was randomly sorted into two more groups: the high-consensus group was told that 81% of their peers agreed with their judgments about race, the low-consensus group was told that 19% of their peers agreed with their judgments about race. After receiving their peers' feedback, participants were asked to wait in a chair in the hallway just outside the experiment room. The hallway was staged with seven chairs, side-by-side. A black research confederate, "who was unaware of the experimental condition of the participants, sat in the seat closest to the door of the experimental room" (ibid., p. 647; see also Figure 4). In short, students had to choose how close to sit to a black peer right after finding out that either most or few of their peers agree with their racial judgments.

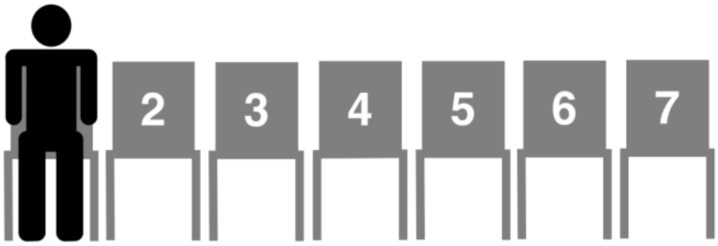

*Figure 4*: Measure of implicit racial bias from Sechrist and Stangor (2001).

Sechrist and Stangor found that highly biased participants in the high-consensus group sat further away from their black peer than their counterparts in the low-consensus group, $F(1, 50) = 5.65$, $p < 0.05$, suggesting that normalizing the biases of high-bias individuals increases their biased behavior. Lowly biased participants in the high-consensus condition sat closer to their black peer than their counterparts in the low-consensus group, $F(1, 50) = 3.22$, $p < 0.07$, suggesting that normalizing the biases of low-bias individuals decreases their biased behavior.

It is no doubt important to investigate whether this peer feedback manipulation is associative or non-associative, reflective or non-reflective. However, even if the nature of this manipulation were discovered, there are a few reasons to resist basing one's view of implicit bias on this particular experiment. First, the analysis might be underpowered, given the sample size of this experiment. A common rule of thumb for sufficient statistical power is to have a minimum of about 50 participants, per experimental condition (Simmons, Nelson, & Simonsohn, 2013, *in press*)—around 4 times the quantity of participants in each of the aforementioned conditions. A proper power analysis could reveal whether this finding is, in fact underpowered, but that would require more information than is reported—e.g., the standard deviations of seating distances in each group. In lieu of a proper power analysis, some researchers recommend estimating power as follows: p = .05 → power ≈ .5; p = .01 → power ≈ .75; p = .005 → power ≈ .8; and p = .001 → power > .9 (Greenwald, Gonzalez, Harris, & Guthrie, 1996). Suffice it to say that the seating distance findings are not well powered, according to this estimation. Worse, recent replication attempts suggest that if this estimation errs, it errs on the side of overestimation (Camerer et al., 2018; Open Science Collaboration, 2015). Power aside, one might still be concerned about the statistical significance of the finding. It is either insignificant or marginally significant even according to the older, lower, and now controversial *p*-value threshold of 0.05 (Benjamin et al., 2017). Also, the reliability of the clever seating measure is not reported. Given the aforementioned psychometric concerns about the IAT, one would want to be reassured about the psychometric validity of this seating measure before accepting the implications of experiments that employ it.

Madva mentions debiasing experiments suggesting that subliminal approach-avoidance behaviors sometimes change implicit racial biases (2017, p. 151). Earlier research had found that repeatedly performing approach behaviors towards subliminal photographs of Black people and avoidance behaviors towards subliminal photographs of White people resulted in a relative preference for the White people on the race IAT, *F*s(2, 41-47) = 2.93-3.18, *p*s = 0.05-0.06 (Kawakami, Phills, Steele, & Dovidio, 2007, 961-966, Experiments 2 and 4). However, more recent research found that this subliminal associative approach and avoidance manipulation did not change race IAT performance compared to controls, *F*(1, 60) = 0.0441, *p* = .83 (Van Dessel, De Houwer, Roets, & Gast, 2016, Experiment 1), suggesting that the earlier manipulations, even if associative, were effective because they were actually *supra*liminal (and therefore, potentially reflective). Indeed, Van Dessel and colleagues claim that this failure to replicate the four earlier findings challenges the idea that implicit biases are "(exclusively) the result of automatic associative learning processes" (Van Dessel, et al., 2016, e2)—notice that their qualificatory use of 'exclusively' avoids the science columnists' any-only mix-up. Nonetheless, Van Dessel and colleagues admit that their null result is "not reliable enough to be treated as conclusive evidence" (ibid., e12) perhaps in part because, as they admit, these studies involve "very small sample[s] of participants" (ibid., e5)—on average, 54 total participants. So, as Van Dessel and colleagues suggest, more research is required to understand whether their manipulations change implicit bias associatively and, orthogonally, whether their manipulations change implicit bias more reflectively.

One might wonder whether the limitations of these experiments apply to other debiasing experiments cited in the debate about the nature of implicit bias. Evaluating the rigor of all relevant debiasing experiments is a worthy inquiry, but it goes beyond the scope of the present paper—see Appendix for additional experiments. The point is just that one should be hesitant to infer a view of implicit bias from the mitigated evidence that is sometimes cited in the debate about the nature of implicit bias.

### 3.3 Strong Evidence

Madva (2017) also mentions more recent, larger, and more methodologically rigorous experiments. One found long-term debiasing while the other found short-term debiasing.

**In-person, long-term Debiasing**. In one of the experiments, 91 non-Black introductory psychology students (67% female, 85% White) were randomly assigned to either a control or an experimental condition after they took the Race IAT (Devine, Forscher, Austin, & Cox, 2012). Both groups completed the Race IAT and typed their results into a computer that explained their results. Then the control group was dismissed but were told that they would need to fill out questionnaires at two points later in the semester. The experimental group was presented with "a 45-minute narrated and interactive slideshow" that educated participants about implicit bias and trained them in five debiasing strategies (ibid., p. 7).

Devine and colleagues found that the experimental manipulation significantly reduced post-test Race IAT results compared to the control groups, $F$ (88) = 7.95, $p$ = 0.006 (Devine et al., 2012, p. 8). And this reduction in implicit bias was maintained (i.e., was not significantly different) 4 and 8 weeks later — $F$ (88) = 0.67, $p$ = 0.42 (ibid.). These findings suggest that certain strategies can manipulate implicitly biased behavior for extended periods of time.

The strategies that Devine and colleagues' participants learned are as follows:

A. *Stereotype replacement*. Identify the stereotypes that inform our responses and replace them with responses that are not based on stereotypes (Monteith, 1993).

B. *Counter-stereotypic imaging*. Imagine counter-stereotypical exemplars when a stereotype is activated (Blair et al., 2001).

C. *Individuation*. Focus on the individual features of someone rather than the stereotypes about them (Brewer, 1988).

D. *Perspective taking*. Imagine the first-person perspective of a member of a stereotyped group rather than the stereotypes about their group (Galinsky & Moskowitz, 2000).

E. *Increasing opportunities for contact*. Seek out positive experiences with members of other groups rather than let oneself imagine stereotypically negative experiences with members of that group (Pettigrew & Tropp, 2006).

Some of these strategies clearly involve conditioning or counterconditioning negative associations—e.g., counter-stereotypic imaging and increasing opportunities for contact (Helton, 2017). Conditioning and counterconditioning are associative manipulations (Mandelbaum 2016, p. 635).

Notice also how much deliberate processing of conscious representations (i.e., reflection) is involved in these strategies: representing a stereotype *as* a stereotype, imagining not just a stereotype but a counter-stereotype, focusing on individual rather than group-level features, imagining the first-person experience of someone with racial features that are different from one's own racial features, and interact positively with people that are negatively stereotyped. So, some of these manipulations involve not only associative processing, but reflective processing.

Nonetheless, some of these debiasing strategies are not known to be purely associative or purely non-associative—e.g., individuation. So, just like in the imagined exploratory experiments (Section 2.2), we are left unable to interpret the roles that associative and non-associative processing play in some of these debiasing strategies.

**Online, short-term debiasing.** In another experiment, around 5000 participants from 17 universities in the United States were randomly assigned to 1 of 9 debiasing conditions or a control condition, after they took the Race IAT (Lai et al., 2016, Study 2). Immediately after completing their condition's requirements, participants took a post-test Race IAT. And two to four days after the experiment, participants completed a second post-test Race IAT.

Lai and colleagues found that eight of the nine debiasing conditions significantly reduced implicitly biased behavior more than the control condition, *F*s(1, 1000-1045) = 6.16-286.73, *p*s = 0.001-0.013 (ibid., pp. 1009-10). Alas, when participants retook the Race IAT two to four days later "[n]one of the interventions had significantly reduced IAT scores relative to control" (Ibid., p. 1010). So, like the previous experiment, the debiasing strategies changed implicitly biased behavior. Yet, unlike the last experiment, the changes did not last.

The debiasing conditions employed by Lai and colleagues debiasing conditions were as follows:

F. *Vivid counterstereotypic scenario*. Read a vivid story about a White villain and a Black hero (Dasgupta & Greenwald, 2001) and keep that in mind during the post-manipulation IAT.

G. *Counterstereotypic IAT*. Practice 32 trials of the IAT in which Black is paired with Good and White is paired with Bad, including some famously positive Black figures such as Oprah and some famously negative White figures such as Hitler (Joy-Gaba & Nosek, 2010).

H. *Competition with shifted group boundaries*. Play a simulated dodgeball game in which one's own teammates are Black and play well and one's opponents are White and play poorly.

I. *Shifting group affiliations under threat*. Read a vivid story about the threat of postnuclear war in which one's closest friends are Black and helpful and one's enemies are White.

J. *Priming multiculturalism*. Read a pro-multiculturalism excerpt, summarize it in one's own words, and list reasons that multiculturalism improves group relations (Richeson & Nussbaum, 2004).

K. *Evaluative conditioning*. Observe 20 Black faces paired with positive words and 20 White faces paired with negative words.

L. *Evaluative conditioning with Go/No-Go task.* Press a button when a Black face is paired with a positive word, do not press a button with a Black face is paired with a negative word, and count the number of Black-positive pairings (Nosek & Banaji, 2001).

M. *Implementation intentions*. Learn that one can override bias by thinking of conditional intentions like, "If I see a Black face, then I will respond by thinking 'good'" (Gollwitzer, 1999).

N. *Faking the IAT*. Learn about Pro-White biases on the IAT and how to intentionally manipulate one's responses times in order for the test to detect a Pro-Black bias (Cvencek, Greenwald, Brown, Gray, & Snowden, 2010).

All conditions except the priming multiculturalism condition clearly involve pairing racial stimuli and valences—i.e., conditioning or counterconditioning. So, at least 8 of Lai and colleagues' manipulations seem to be associative (Mandelbaum 2016, p. 635). Whether the priming multiculturalism condition counts as purely non-associative will be controversial given the disagreement about whether so-called logical and evidential manipulations involve associative processing (e.g., Briñol et al., 2009 vs. Mandelbaum, 2016). Until such disagreement is resolved or at least uncontroversial and until it is clear that priming multiculturalism can produce lasting changes in implicit bias, experiments that prime multiculturalism are exploratory debiasing experiments (Section 2.2) and therefore unable to determine whether implicit bias is associative or non-associative.

Notice also that some of Lai and colleagues' associative manipulations seem to involve deliberate processing of conscious representations—e.g., pre-emptively thinking "Black = Good" (2016, pp. 1005). This suggests that some of Lai and colleagues' associative manipulations involved reflection.

**Duration of experiment & debiasing.** The evidence for long-term debiasing was mixed. Devine and colleagues found that counterconditioning-like protocols produced short- and long-term changes in implicitly biased behavior, but Lai and colleagues found that counterconditioning resulted in only short-term changes. Three details about this mixed evidence are worth emphasizing.

First, consider the conflict about long-term findings. Given that Lai and colleagues' sample was much larger and was collected from multiple locations, its population is more representative, and its analysis confers greater statistical power. So, if one had to bet on the likelihood of long-term debiasing via counterconditioning across populations, then one should bet against them—until further debiasing experiments suggest otherwise, of course. Nonetheless, we might wonder if this conflict is only apparent. That is, perhaps long-term debiasing works in certain subsets of the population—like Devine and colleagues' non-Black, mostly-White undergraduate sample—even if long-term debiasing does not work, on average, across the population as a whole. Of course, this is an empirical hypothesis.

Second, the conflict about long-term findings might be related to the duration, frequency, or even context of the counterconditioning manipulations. Notably, previous work found that 5 minutes of counterconditioning in a controlled setting did not significantly change implicitly biased behavior, but four blocks of 96 trials of counterconditioning did (Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000). Similarly, Lai and colleagues found that short, online debiasing protocols did not lead to long-term changes, but Devine and colleagues found that teaching students how debiasing works in everyday social settings did. Taken together, one might hypothesize that long-term debiasing is more likely with further counterconditioning (Van Dessel, De Houwer, Roets, & Gast, 2016, e12)—and not just in lab settings, but in everyday social settings (Madva, 2017).

Third, notice a consistency between Devine and colleagues' and Lai and colleagues' findings: more or less reflective counterconditioning changed implicitly biased behavior, even if only briefly. This finding is also consistent with earlier work (e.g., Olson & Fazio, 2006, Experiment 2; Rydell & McConnell, 2006).

## 4 CRITICAL DISCUSSION

With the inferential principles and the evidence on the table, we are now prepared to determine whether the received view of implicit bias should be abandoned for more centrist or far-right views (Table 2). I will argue that the received, unreflective, associationist view of implicit bias should be abandoned for a more general associationist view of implicit bias, given the strong evidence just considered. However, I will concede that if certain evidence exists, now or in the future, even this more general associationist view of implicit bias should be abandoned for either interactionism about implicit bias or minimalism about implicit bias—views that will be explained below.

### 4.1 Associationism & Reflectivism About Implicit Bias

The strongest evidence found that conditioning or counterconditioning can change implicitly biased behavior—even if only briefly. Conditioning and counterconditioning are widely accepted to be associative manipulations (Mandelbaum 2016, p. 635).

**Associationism.** Via the affirmative manipulation principle, one can infer that implicit bias is at least partly associative. It may be tempting to conclude from this that associationism about implicit bias is true and, therefore, that non-associationism about implicit bias is false. However, that relies on the problematic negative manipulation principle that leads to the science columnists' any-only mix-up: in this case, the mistake of concluding that implicit bias is not predicated on *any* non-associative processes when the evidence merely shows that something is not predicated on *only* non-associative processes.

**Reflectivism.** To test for positive evidence that implicit bias is also partly non-associative, some debiasing experiments dissociate the effect of associative processing on implicit bias from the effect(s) of other kinds of processing (e.g., Calanchini, Gonsalkorale, Sherman, & Klauer, 2013). These process dissociation debiasing experiments find that debiasing is explained by both (a) the degree to which associations are activated and (b) the degree to which participants reflect on appropriate responses. The fact that reflection can change implicitly biased behavior is consistent with the strongest evidence under consideration. If reflection were necessarily non-associative, then the negative intervention principle would allow us to infer from this evidence that implicit bias is not predicated on only associative processes. However, many have realized that reflection is not necessarily non-associative (Section 1).

Nonetheless, the fact that reflection can help reduce implicitly biased behavior supports a sort of reflectivism about implicit bias. Reflectivism is just the idea that reflection is an important part of improving our judgments and behavior (Doris, 2015; Ferrin, 2017). And reflection seems to be involved in counterconditioning implicit bias. This is not to say that there is strong evidence for an infallibilist reflectivism, according to which reflection fully or permanently ameliorates implicit bias. Rather, the strong evidence suggests only a kind of "sensible reflectivism" according to which reflection can—but does not necessarily—ameliorate implicit bias, albeit only briefly and incompletely (Schwenkler, 2018). If that is right, then we can infer, via the affirmative manipulation principle that implicit bias can be reflective.

### 4.2 Interactionism & Minimalism About Implicit Bias

This paper has focused on implicit *racial* bias and on three categories of evidence from experimental attempts to reduce such biases. Given how many debiasing experiments have been conducted (e.g., see Appendix) and how thoroughly one should analyze these experiments, one paper cannot sufficiently review all racial debiasing experiments—let alone all debiasing experiments. So, there may be strong evidence, now or in the future, of non-associative interventions on implicitly biased behavior. If or when such evidence exists, then associationism about implicit bias would be false: i.e., implicitly biased behavior would not be predicated on only associative processing.

Of course, falsifying associationism about implicit bias would not support non-associationism about implicit bias, given the strong evidence already considered. That is, if we add non-associative interventions on implicitly biased behavior to our body of strong evidence, then our total evidence would suggest that implicit bias can be changed via associative processes, given the strong evidence considered herein, as well as non-associative processes, given the additional evidence. According to the affirmative manipulation principle, that total evidence entails that implicit bias can be predicated on either associative or non-associative processes. Such a disjunctive conclusion brings us to a fork in the road.

**Interactionism.** Down one side of the fork, there are interactionist views of implicit bias. Interactionist views of implicit bias accept that implicit bias is predicated on associative and non-associative processes. However, interactionist views also aim to describe precisely *how* these processes interact to produce the observed dynamics in implicit biases (e.g., by testing for manipulation patterns matching 3c, 3d, 3e, and 3f in Figure 3). In other words, the goal of an interactionist view is its attempt to construct and test cognitive models of implicit bias. There are a variety of interactionist views that seem to accomplish this goal (e.g., Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Gawronski & Bodenhausen, 2014; Perugini, 2005).

**Minimalism.** Interactionist views of implicit bias include more cognitive details than are necessary for some philosophers' claims about implicit bias. So, some philosophers can go down the other side of the fork toward minimalist views of implicit bias. Minimalist views accept the complexity of implicit bias. They acknowledge that implicit bias can seem associative in some circumstances and seem non-associative in other circumstances. Minimalist views of implicit bias also acknowledge that implicit bias seems to involve more reflection in some circumstances and less reflection in other circumstances. Crucially, however, minimalist views of implicit bias do not aim to provide a falsifiable account of whether and how implicit bias is predicated on certain types of cognitive processing. Rather, the goal of minimalism about implicit bias is to account for our normative intuitions about cases of implicit bias without relying on any particular cognitive model of implicit bias. There are various discussions of implicit bias that might be able to accomplish these goals (e.g., Levy, 2016; Smith *forthcoming*; Sullivan-Bissett, 2015).

Of course, the goals of minimalist views and interactionist views are not mutually exclusive. After all, while some normative intuitions about implicit bias need not commit to any particular cognitive model of implicit bias, some cognitive models of implicit bias, if justified, will justify some normative intuitions about implicit bias more than others (e.g., Huebner 2016 and Toribio, 2018a). So, there are advantages to basing normative claims about implicit bias on the most promising interactionist views of implicit bias. And some philosophers seem to realize as much (e.g., Berger, 2018; Holroyd & Sweetman, 2016; Madva, 2017; Levy, 2015).

## 5 CONCLUSION

On the one hand, the conclusion of this investigation is somewhat progressive. The stronger debiasing evidence under consideration did not support the received, unreflective, associationist views of implicit bias. On the contrary, there was strong evidence that implicitly biased behavior can also be changed via more reflective processing, supporting the more capacious associationist views of implicit bias. On the other hand, the conclusion of this investigation is somewhat conservative. While the received view of implicit bias was only partially supported by strong debiasing experiments, the road to far-right, non-associationist views of implicit bias remained blocked by strong evidence of associative debiasing manipulations.

Nonetheless, future or overlooked evidence might clearly show that non-associative manipulations change implicitly biased behavior. If that is the case, then more centrist, interactionist views of implicit bias can be inferred. Otherwise, two views of implicit bias can be inferred from debiasing experiments: first, associationist views of implicit bias and second, minimalist views of implicit bias.

Of course, there are limitations to the existing investigation. First, views about implicit bias are based on more than just debiasing experiments. So, the conclusions about implicit bias that were inferred from debiasing experiments herein are not all-things-considered conclusions. As such, views of implicit bias may be further supported or further undermined by considerations besides debiasing experiments. Second, the debate between associationism or non-associationism about implicit bias is just a specific instance of the more general debate between associationism or non-associationism about mind. So, this investigation cannot settle that general debate. However, this investigation can recommend that debaters avoid the science columnists' any-only mix-up by relying on the negative intervention principle rather than the negative manipulation principle.

## REFERENCES

Anderson, J. R., & Bower, G. H. (1980). Human Associative Memory. New Jersey: Lawrence Erlbaum Associates, Inc.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109. doi:10.1016/j.cognition.2016.10.014

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, 46(3), 668–688. doi:10.3758/s13428-013-0410-6

Bargh, J. A. (1992). The Ecology of Automaticity: Toward Establishing the Conditions Needed to Produce Automatic Processing Effects. The American Journal of Psychology, 105(2), 181–199. doi:10.2307/1423027

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2017). Redefine statistical significance. Nature Human Behaviour, 1.

Berger, J. (2018). Implicit attitudes and awareness. *Synthese*, 1–22. doi:10.1007/s11229-018-1754-3

Blake, A. (2016, September 26). The first Trump-Clinton presidential debate transcript, annotated. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-fix/wp/2016/09/26/the-first-trump-clinton-presidential-debate-transcript-annotated/

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. Journal of Personality and Social Psychology, 81(5), 828.

Brewer, M. B. (1988). A dual process model of impression formation. Advances in Social Cognition, 1.

Briñol, P., Petty, R. E., & McCaslin, M. J. (2009). Changing attitudes on implicit versus explicit measures: What is the difference. In R. H. Fazio, R. E. Petty, & P. Briñol (Eds.), Attitudes: Insights from the new implicit measures (pp. 285–326). Psychology Press.

Brownstein, M. (2018). *The Implicit Mind: Cognitive Architecture, the Self, and Ethics*. New York, NY: Oxford University Press.

Calanchini, J., Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2013). Counter-prejudicial training reduces activation of biased associations and enhances response monitoring. *European Journal of Social Psychology*, *43*(5), 321–325. doi:10.1002/ejsp.1941

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 1. doi:10.1038/s41562-018-0399-z

Carley, L. (2018, October 31). Breaking the Bias Habit: An Evidence-Based Intervention in Duke's Biology Department | Duke Graduate School. Retrieved November 6, 2018, from https://gradschool.duke.edu/professional-development/blog/breaking-bias-habit-evidence-based-intervention-duke-s-biology

Carnap, R. (1950). Logical Foundations of Probability. Chicago: University of Chicago Press.

Cone, J., Mann, T. C., & Ferguson, M. J. (2017). Chapter Three - Changing Our Implicit Minds: How, When, and Why Implicit Evaluations Can Be Rapidly Revised. In J. M. Olson (Ed.), *Advances in Experimental Social Psychology* (Vol. 56, pp. 131–199). Academic Press. doi:10.1016/bs.aesp.2017.03.001

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating Multiple Processes in Implicit Social Cognition: The Quad Model of Implicit Task Performance. *Journal of Personality and Social Psychology*, *89*(4), 469–487. doi:10.1037/0022-3514.89.4.469

Corneille, O., & Stahl, C. (2018). Associative Attitude Learning: A Closer Look at Evidence and How It Relates to Attitude Models. *Personality and Social Psychology Review*, 1088868318763261. doi:10.1177/1088868318763261

Cvencek, D., Greenwald, A. G., Brown, A. S., Gray, N. S., & Snowden, R. J. (2010). Faking of the Implicit Association Test Is Statistically Detectable and Partly Correctable. *Basic and Applied Social Psychology*, *32*(4), 302–314. doi:10.1080/01973533.2010.519236

Dacey, M. (2016). Rethinking associations in psychology. Synthese, 1–24.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, *81*(5), 800–814. doi:10.1037//0022-3514.81.5.800

De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, 37(2), 176–187. Doi:10.1016/j.lmot.2005.12.002

De Houwer, J. (2018). Propositional Models of Evaluative Conditioning. *Social Psychological Bulletin*, *13(3)*, e28046. doi:10.5964/spb.v13i3.28046

Del Pinal, G. D., & Spaulding, S. (2018). Conceptual centrality and implicit bias. *Mind & Language*, *33*(1), 95–111. doi:10.1111/mila.12166

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. Journal of Experimental Social Psychology, 48(6), 1267–1278.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5–18. doi:10.1037/0022-3514.56.1.5

Doris, J. M. (2015). *Talking to Our Selves: Reflection, Ignorance, and Agency*. OUP Oxford.

Evans, J. S. B. T. (2009). How Many Dual Process Theories Do We Need: One, Two or Many? In J. S. B. T. Evans & K. Frankish (Eds.), *In Two Minds: Dual Processes and Beyond* (pp. 31–54). Oxford: Oxford University Press.

Evans, J., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition Advancing the Debate. Perspectives on Psychological Science, 8(3), 223–241.

Ferrin, A. (2017). Good Moral Judgment and Decision-Making Without Deliberation. *The Southern Journal of Philosophy*, *55*(1), 68–95. doi:10.1111/sjp.12210

Forscher, P. S., Lai, C., Axt, J., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2018). A Meta-Analysis of Procedures to Change Implicit Measures. doi:10.31234/osf.io/dv8tu

Frankish, K. (2010). Dual-Process and Dual-System Theories of Reasoning. Philosophy Compass, 5(10), 914–926.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. doi:10.1257/089533005775196732

Fridland, E. (2016). Skill and motor control: intelligence all the way down. Philosophical Studies, 174(6), 1539–1560. doi:10.1007/s11098-016-0771-7

———. (*forthcoming*). Longer, Smaller, Faster, Stronger, On Skills and Intelligence. Philosophical Psychology.

Gaertner, S. L., & McLaughlin, J. P. (1983). Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics. *Social Psychology Quarterly*, *46*(1), 23–30. doi:10.2307/3033657

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. Journal of Personality and Social Psychology, 78(4), 708.

Gawronski, B. (forthcoming). Six Lessons for a Cogent Science of Implicit Bias and Its Criticism. *Perspectives on Psychological Science*. Retrieved from https://www.researchgate.net/publication/329656554_Six_Lessons_for_a_Cogent_Science_of_Implicit_Bias_and_Its_Criticism

Gawronski, B., & Bodenhausen, G. V. (2014). The associative-propositional evaluation model: Operating principles and operating conditions of evaluation. *Dual-Process Theories of the Social Mind*, 188–203.

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition*, 15(3), 485–499. doi:10.1016/j.concog.2005.11.007

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis. *Personality and Social Psychology Bulletin*, 43(3), 300–312. doi:10.1177/0146167216684131

Gawronski, B., Walther, E., & Blank, H. (2005). Cognitive consistency and the formation of interpersonal attitudes: Cognitive balance affects the encoding of social information. Journal of Experimental Social Psychology, 41(6), 618–626. doi:10.1016/j.jesp.2004.10.005

Gendler, T. S. (2008a). Alief and Belief. The Journal of Philosophy, 105(10), 634–663.

———. (2008b). Alief in action (and reaction). Mind & Language, 23(5), 552–585.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*(7), 493–503.

Greenwald, A. G., Andrew, T., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17–41. doi:10.1037/a0015575

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. Journal of Personality and Social Psychology, 108(4), 553–561. doi:10.1037/pspa0000016

Greenwald, A., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology*, *33*(2), 175–183. doi:10.1111/j.1469-8986.1996.tb02121.x

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology, 74(6), 1464.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes., Awareness of Implicit Attitudes. *Journal of Experimental Psychology. General, Journal of Experimental Psychology. General*, *143*, *143*(3, 3), 1369, 1369–1392. doi:10.1037/a0035028, 10.1037/a0035028

Helton, G. (2017, March 23). Personal Communication at 109th Annual Meeting of the Southern Society for Psychology and Philosophy.

Holroyd, J., & Sweetman, J. (2016). The Heterogeneity of Implicit Bias. In M. Brownstein & J. Saul (Eds.), *Implicit Bias and Philosophy*. Oxford University Press.

Huebner, B. (2016). Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition. In M. Brownstein & J. Saul (Eds.), Implicit Bias and Philosophy, Vol 1. Oxford University Press.

Hume, D. (1978). *A Treatise of Human Nature*. (L. A. Selby-Bigge & P. H. Nidditch, Eds.) (2nd edition). Oxford; New York: Oxford University Press.

———. (1983). *An Enquiry Concerning the Principles of Morals*. (E. Steinberg & J. B. Schneewind, Eds.). Indianapolis: Hackett Publishing.

Hütter, M., & Sweldens, S. (2018). Dissociating Controllable and Uncontrollable Effects of Affective Stimuli on Attitudes and Consumption. *Journal of Consumer Research*, *45*(2), 320–349. doi:10.1093/jcr/ucx124

Jost, J. T. (2018). The IAT Is Dead, Long Live the IAT: Context-Sensitive Measures of Implicit Attitudes Are Indispensable to Social and Political Psychology. *Current Directions in Psychological Science*, 0963721418797309. doi:10.1177/0963721418797309

Joy-Gaba, J. A., & Nosek, B. A. (2010). The Surprisingly Limited Malleability of Implicit Racial Evaluations. *Social Psychology*, *41*(3), 137–146. doi:10.1027/1864-9335/a000020

Kahneman, D. (2011). Thinking, Fast and Slow. Macmillan.

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, *78*(5), 871–888.

Korsgaard, C. M. (1996). *The Sources of Normativity*. Cambridge University Press.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. Nature Communications, 6, 7455.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. Journal of Experimental Psychology. General, 145(8), 1001–1016.

Levy, N. (2015). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Noûs*, 49(4), 800–823.

———. (2016). Implicit Bias and Moral Responsibility: Probing the Data. Philosophy and Phenomenological Research, 93(3).

Machery, E. (2016). De-Freuding implicit attitudes. In M. Brownstein & J. Saul (Eds.), Implicit Bias and Philosophy, Vol 1 (Vol. 1, pp. 104–129). Oxford University Press, Vol 1.

Madva, A. (2015). Why implicit attitudes are (probably) not beliefs. *Synthese*, *193*(8), 2659–2684. doi:10.1007/s11229-015-0874-2

———. (2017). Biased against Debiasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle against Prejudice. *Ergo, an Open Access Journal of Philosophy*, *4*. doi:10.3998/ergo.12405314.0004.006

Mandelbaum, E. (2017). Associationist Theories of Thought. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2017). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2017/entries/associationist-thought/

———. (2016). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. Noûs, 50(3), 629-

——— (2013). Thinking is Believing. Inquiry, 57(1), 55–96.

McCloskey, D. N., & Ziliak, S. T. (1996). The Standard Error of Regressions. *Journal of Economic Literature*, *34*(1), 97–114.

McCoy, M. K. (2018, June 1). Researcher: Despite Good Intentions, Anti-Bias Training Can Actually Backfire. *Wisconsin Public Radio*. Retrieved from https://www.wpr.org/researcher-despite-good-intentions-anti-bias-training-can-actually-backfire

Melnikoff, D. E., & Bargh, J. A. (2018). The Mythical Number Two. *Trends in Cognitive Sciences*, *22*(4), 280–293. doi:10.1016/j.tics.2018.02.001

Meyer, D. (2018, May 29). Starbucks Is Closing Today For Its Company-Wide Unconscious Bias Training: Here's What You Need To Know. *Fortune*. Retrieved from http://fortune.com/2018/05/29/starbucks-closing-today-unconscious-bias-training/

Monteith, M. J. (1993). Self-regulation of prejudiced responses: Implications for progress in prejudice-reduction efforts. Journal of Personality and Social Psychology, 65(3), 469.

Moors, A., & De Houwer, J. (2006). Automaticity: A Theoretical and Conceptual Analysis. Psychological Bulletin, 132(2), 297–326. doi:10.1037/0033-2909.132.2.297

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. Computer Science Department, Paper 2033.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition*, *19*(6), 625–666. doi:10.1521/soco.19.6.625.20886

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. doi:10.1126/science.aac4716

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, *108*(4), 562–571. doi:10.1037/pspa0000023

Payne, K., Niemi, L., & Doris, J. (2018, March 27). How to Think about "Implicit Bias." *Scientific American*. Retrieved from https://www.scientificamerican.com/article/how-to-think-about-implicit-bias/

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2015). Is the cognitive reflection test a measure of both reflection and intuition? Behavior Research Methods, 1–8.

Pennycook, G., Fugelsang, J. A., Koehler, D. J., & Thompson, V. A. (2016). Commentary: Rethinking fast and slow based on a critique of reaction-time reverse inference. Frontiers in Psychology, 7.

Pennycook, G., Neys, W. D., Evans, J. S. B. T., Stanovich, K. E., & Thompson, V. A. (2018). The Mythical Dual-Process Typology. *Trends in Cognitive Sciences*. doi:10.1016/j.tics.2018.04.008

Perugini, M. (2005). Predictive models of implicit and explicit attitudes. *British Journal of Social Psychology*, *44*(1), 29–45. doi:10.1348/014466604X23491

Perugini, M., Richetin, J., & Zogmaister, C. (2010). Prediction of behavior. *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*, *10*, 255–278.

Peters, U. (*forthcoming*). Implicit bias, ideological bias, and epistemic risks in philosophy. *Mind & Language*, *0*(0). doi:10.1111/mila.12194

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. Journal of Personality and Social Psychology, 90(5), 751.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and Cognitive Control. In R. L. Solso (Ed.), Information Processing and Cognition: The Loyola Symposium. Lawrence Erlbaum.

Quilty-Dunn, J., & Mandelbaum, E. (2017). Against dispositionalism: belief in cognitive science. Philosophical Studies, 1–20. doi:10.1007/s11098-017-0962-x

Richeson, J. A., & Nussbaum, R. J. (2004). The impact of multiculturalism versus color-blindness on racial bias. *Journal of Experimental Social Psychology*, *40*(3), 417–423. doi:10.1016/j.jesp.2003.09.002

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. Journal of Personality and Social Psychology, 91(6), 995–1008. doi:10.1037/0022-3514.91.6.995

Saul, J. (2013a). Implicit bias, stereotype threat and women in philosophy. In K. Hutchison & F. Jenkins (Eds.), *Women in philosophy: What needs to change* (pp. 39–60). Oxford University Press.

Saul, J. (2013b). Scepticism and Implicit Bias. *Disputatio*, 5(37), 243–263.

Schwenkler, J. (2018). Self-Knowledge and Its Limits. *Journal of Moral Philosophy*, *15*(1), 85–95. doi:10.1163/17455243-01501005

Schwitzgebel, E. (2002). A Phenomenal, Dispositional Account of Belief. Noûs, 36(2), 249–275.

———. (2010). Acting Contrary to Our Professed Beliefs or the Gulf Between Occurrent Judgment and Dispositional Belief. Pacific Philosophical Quarterly, 91(4), 531–553.

Sechrist, G. B., & Stangor, C. (2001). Perceived consensus influences intergroup behavior and stereotype accessibility. Journal of Personality and Social Psychology, 80(4), 645–654.

Shea, N., & Frith, C. D. (2016). Dual-process theories and consciousness: the case for 'Type Zero' cognition. Neuroscience of Consciousness, 2016(1).

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (in press). False-Positive Citations. Perspectives on Psychological Science.

———. (2013). Life after P-Hacking. In Meeting of the Society for Personality and Social Psychology (p. 38). New Orleans, LA.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 119(1), 3–22. doi:10.1037/0033-2909.119.1.3

Strack, F., & Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior. Personality and Social Psychology Review, 8(3), 220–247.

Smith, A. (forthcoming). Implicit Bias, Moral Agency, and Moral Responsibility. In *The Norton Introduction to Philosophy*. New York: W. W. Norton & Company.

Sullivan-Bissett, E. (2015). Implicit bias, confabulation, and epistemic innocence. *Consciousness and Cognition*, *33*, 548–560. doi:10.1016/j.concog.2014.10.006

Sun, R. (2016). Implicit and Explicit Processes: Their Relation, Interaction, and Competition. In L. Macchi, M. Bagassi, & R. Viale (Eds.), *Cognitive Unconscious and Human Rationality* (pp. 257–274). Cambridge, MA: The MIT Press.

Toribio, J. (2018a). Accessibility, implicit bias, and epistemic justification. *Synthese*, 1–19. doi:10.1007/s11229-018-1795-7

———. (2018b). Implicit Bias: From Social Structure to Representational Format. *Theoria: An International Journal for Theory, History and Foundations of Science*, *33*(1), 41–60.

Van Dessel, P., De Houwer, J., Roets, A., & Gast, A. (2016). Failures to change stimulus evaluations by means of subliminal approach and avoidance training. *Journal of Personality and Social Psychology*, *110*(1), e1–e15. doi:10.1037/pspa0000039

Woodward, J. (2016). Causation and Manipulability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2016/entries/causation-mani/

The following list includes additional experiments that attempt to change implicit biases in behavior. These experiments involve instances of implicit bias than could not be discussed in sufficient detail in the main text, such as implicit biases about gender, sexual orientation, political orientation, consumer products, substance use, pseudowords, and more. Some of these experiments employ indirect measures of bias besides the IAT. The list was composed of recommendations from reviewers, various conference participants, and Google Scholar alerts for new publications by or related to authors already cited in the main text.

Andreychik, M. R., & Gill, M. J. (2012). Do negative implicit associations indicate negative attitudes? Social explanations moderate whether ostensible "negative" associations are prejudice-based or empathy-based. *Journal of Experimental Social Psychology*, 48(5), 1082–1093. Doi:10.1016/j.jesp.2012.05.006

Arendt, F., Marquart, F., & Matthes, J. (2015). Effects of Right-Wing Populist Political Advertising on Implicit and Explicit Stereotypes. *Journal of Media Psychology*, 1–12. Doi:10.1027/1864-1105/a000139

Arendt, F., & Northup, T. (2015). Effects of Long-Term Exposure to News Stereotypes on Implicit and Explicit Attitudes. *International Journal of Communication*, 9(0), 21.

Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability From 2007 to 2016. *Psychological Science*, 0956797618813087. Doi:10.1177/0956797618813087

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic Preference for White Americans: Eliminating the Familiarity Explanation. *Journal of Experimental Social Psychology*, 36(3), 316–328. Doi:10.1006/jesp.1999.1418

Dasgupta, N., & Rivera, L. M. (2008). When Social Context Matters: The Influence of Long–Term Contact and Short–Term Exposure to Admired Outgroup Members on Implicit Attitudes and Behavioral Intentions. *Social Cognition*, 26(1), 112–123. Doi:10.1521/soco.2008.26.1.112

Everett, J. A. C., Schellhaas, F. M. H., Earp, B. D., Ando, V., Memarzia, J., Parise, C. V., … Hewstone, M. (2014). Covered in stigma? The impact of differing levels of Islamic head-covering on explicit and implicit biases toward Muslim women. Journal of Applied Social Psychology, 45(2), 90–104. Doi:10.1111/jasp.12278

Forehand, M. R., & Perkins, A. (2005). Implicit Assimilation and Explicit Contrast: A Set/Reset Model of Response to Celebrity Voice-*Overs. Journal of Consumer Research*, 32(3), 435–441. Doi:10.1086/497555

French, A. R., Franz, T. M., Phelan, L. L., & Blaine, B. E. (2013). Reducing Muslim/Arab Stereotypes Through Evaluative Conditioning. *The Journal of Social Psychology*, 153(1), 6–9. Doi:10.1080/00224545.2012.706242

Gawronski, B., Bodenhausen, G. V., & Becker, A. P. (2007). I like it, because I like myself: Associative self-anchoring and post-decisional change of implicit evaluations. *Journal of Experimental Social Psychology*, 43(2), 221–232. Doi:10.1016/j.jesp.2006.04.001

Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology*, 40(4), 535–542. Doi:10.1016/j.jesp.2003.10.005

Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology*, 44(5), 1355–1361. Doi:10.1016/j.jesp.2008.04.005

Han, H. A., Olson, M. A., & Fazio, R. H. (2006). The influence of experimentally created extrapersonal associations on the Implicit Association Test. *Journal of Experimental Social Psychology*, 42(3), 259–272. Doi:10.1016/j.jesp.2005.04.006

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, 136(3), 390–421. Doi:10.1037/a0018916

Horcajo, J., Briñol, P., & Petty, R. E. (2010). Consumer persuasion: Indirect change and implicit balance. *Psychology & Marketing*, 27(10), 938–963. Doi:10.1002/mar.20367

Hu, X., Antony, J. W., Creery, J. D., Vargas, I. M., Bodenhausen, G. V., & Paller, K. A. (2015). Unlearning implicit social biases during sleep. *Science*, 348(6238), 1013–1015. Doi:10.1126/science.aaa3841

Hughes, S., Ye, Y., Van Dessel, P., & De Houwer, J. (2018). When People Co-occur With Good or Bad Events: Graded Effects of Relational Qualifiers on Evaluative Conditioning. *Personality and Social Psychology Bulletin*, 0146167218781340. Doi:10.1177/0146167218781340

Kinoshita, S., & Peek-O'leary, M. (2006). Two bases of the compatibility effect in the Implicit Association Test (IAT). *The Quarterly Journal of Experimental Psychology*, 59(12), 2102–2120. Doi:10.1080/17470210500451141

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., … Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology. General*, 143(4), 1765–1785. Doi:10.1037/a0036260

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions?: The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology*, 108(6), 823–849. Doi:10.1037/pspa0000021

Mele, M. L., Federici, S., & Dennis, J. L. (2014). Believing Is Seeing: Fixation Duration Predicts Implicit Negative Attitudes. *PLoS ONE*, 9(8), e105106. Doi:0.1371/journal.pone.0105106

Miles, E., & Crisp, R. J. (2014). A meta-analytic test of the imagined contact hypothesis. *Group Processes & Intergroup Relations*, 17(1), 3–26. Doi:10.1177/1368430213510573

Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious Control of Stereotype Activation Through Chronic Egalitarian Goals. *Journal of Personality and Social Psychology*, 77(1), 167–184. Doi:10.1037//0022-3514.77.1.167

Ramos, M. R., Barreto, M., Ellemers, N., Moya, M., Ferreira, L., & Calanchini, J. (2016). Exposure to sexism can decrease implicit gender stereotype bias. *European Journal of Social Psychology*, 46(4), 455–466. Doi:10.1002/ejsp.2165

Rothermund, K., & Wentura, D. (2004). Underlying Processes in the Implicit Association Test: Dissociating Salience From Associations. *Journal of Experimental Psychology: General*, 133(2), 139–165. Doi:10.1037/0096-3445.133.2.139

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). Unlearning automatic biases: the malleability of implicit prejudice and stereotypes. *Journal of Personality and Social Psychology*, 81(5), 856–868.

Rudman, L. A., & Lee, M. R. (2002). Implicit and Explicit Consequences of Exposure to Violent and Misogynous Rap Music. *Group Processes & Intergroup Relations*, 5(2), 133–150. Doi:10.1177/1368430202005002541

Rudman, L. A., & Phelan, J. E. (2010). The effect of priming gender roles on women's implicit gender beliefs and career aspirations. *Social Psychology*, 41(3), 192–202. Doi:10.1027/1864-9335/a000027

Sellaro, R., Derks, B., Nitsche, M. A., Hommel, B., van den Wildenberg, W. P. M., van Dam, K., & Colzato, L. S. (2015). Reducing Prejudice Through Brain Stimulation. *Brain Stimulation*, 8(5), 91–897. Doi:10.1016/j.brs.2015.04.003

Smith, C. T., & De Houwer, J. (2015). Hooked on a feeling: affective anti-smoking messages are more effective than cognitive messages at changing implicit evaluations of smoking. *Frontiers in Psychology*, 6. Doi:10.3389/fpsyg.2015.01488

Smith, C. T., De Houwer, J., & Nosek, B. A. (2013). Consider the Source: Persuasion of Implicit Evaluations Is Moderated by Source Credibility. *Personality and Social Psychology Bulletin*, 39(2), 193–205. Doi:10.1177/0146167212472374

Stell, A. J., & Farsides, T. (2015). Brief loving-kindness meditation reduces racial bias, mediated by positive other-regarding emotions. *Motivation and Emotion*, 1–8. Doi:10.1007/s11031-015-9514-x

Stewart, T. L., Latu, I. M., Kawakami, K., & Myers, A. C. (2010). Consider the situation: Reducing automatic stereotyping through Situational Attribution Training. *Journal of Experimental Social Psychology*, 46(1), 221–225. Doi:10.1016/j.jesp.2009.09.004

Tello, N., Bocage-Barthélémy, Y., Dandaba, M., Jaafari, N., & Chatard, A. (2018). Evaluative conditioning: A brief computer-delivered intervention to reduce college student drinking. *Addictive Behaviors*, 82, 14–18. Doi:10.1016/j.addbeh.2018.02.018

Vanaelst, J., Spruyt, A., & De Houwer, J. (2016). How to Modify (Implicit) Evaluations of Fear-Related Stimuli: Effects of Feature-Specific Attention Allocation. *Psychopathology*, 717. Doi:10.3389/fpsyg.2016.00717

Van Dessel, P., Ye, Y., & De Houwer, J. (2018). Changing Deep-Rooted Implicit Evaluation in the Blink of an Eye: Negative Verbal Information Shifts Automatic Liking of Gandhi. *Social Psychological and Personality Science*, 1948550617752064. Doi:10.1177/1948550617752064

Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2017). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology*, 69, 23-32.

Van Dessel, P., De Houwer, J., & Smith, C. T. (2017). Relational information moderates approach-avoidance instruction effects on implicit evaluation. *Acta Psychologica*. Doi:10.1016/j.actpsy.2017.03.016

Van Dessel, P., De Houwer, J., Roets, A., & Gast, A. (2016). Failures to change stimulus evaluations by means of subliminal approach and avoidance training. *Journal of Personality and Social Psychology*, 110(1), e1-e15. Doi:10.1037/pspa0000039

Vorauer, J. D. (2012). Completing the implicit association test reduces positive intergroup interaction behavior. *Psychological Science*, 23(10), 1168–1175. Doi:10.1177/0956797612440457

Waiguny, M. K. J., Nelson, M. R., & Marko, B. (2013). How Advergame Content Influences Explicit and Implicit Brand Attitudes: When Violence Spills Over. *Journal of Advertising*, 42(2–3), 155–169. Doi:10.1080/00913367.2013.774590

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for Racial Prejudice at the Implicit Level and Its Relationship With Questionnaire Measures. *Journal of Personality and Social Psychology*, 72(2), 262–274.

Yoshida, E., Peach, J. M., Zanna, M. P., & Spencer, S. J. (2012). Not all automatic associations are created equal: How implicit normative evaluations are distinct from implicit attitudes and uniquely predict meaningful behavior. *Journal of Experimental Social Psychology*, 48(3), 694–706. Doi:10.1016/j.jesp.2011.09.013

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning? *Quarterly Journal of Experimental Psychology*, 67(11), 2105-2122. Doi:10.1080/17470218.2014.907324