Synthese
https://doi.org/10.1007/s11229-021-03401-3

ORIGINAL RESEARCH

'I *love* women': an explicit explanation of implicit bias test results

Samuel Reis-Dennis[1] · **Vida Yao**[2]

Received: 6 April 2021 / Accepted: 2 September 2021

Abstract
Recent years have seen a surge of interest in implicit bias. Driving this concern is the thesis, apparently established by tests such as the IAT, that people who hold egalitarian explicit attitudes and beliefs are often influenced by implicit mental processes that operate independently from, and are largely insensitive to, their explicit attitudes. We argue that implicit bias testing in social and empirical psychology does not, and without a fundamental shift in focus could not, establish this startling thesis. We suggest that implicit bias research has been conducted in light of inadequate theories of racism and sexism. As a result, such testing has not sufficiently controlled for subjects' prejudiced explicit beliefs and emotions, and has not ruled out the possibility that explicit prejudice best explains test subjects' discriminatory associations and behavior.

Keywords Implicit bias · Racism · Sexism · Moral Psychology · Responsibility

1 Introduction

The concept of implicit bias has generated intense interest among philosophers and the public. Yet the empirical research upon which this enthusiasm is founded has recently come under scrutiny. Critics have worried that implicit bias tests do not reliably predict subjects' performance of prejudiced actions better than explicit bias tests; they have emphasized the conflicting and anomalous data such tests yield; and they have argued

✉ Samuel Reis-Dennis
reisdes@amc.edu

Vida Yao
vy2@rice.edu

[1] Alden March Bioethics Institute, Albany Medical College, 47 New Scotland Avenue MC 158, Albany, NY 12208, USA

[2] Department of Philosophy, Rice University, 6100 Main St., Houston, TX 77005, USA

Published online: 24 September 2021

that academic and political focus on implicit explanations for prejudice draws attention and resources away from efforts to understand and mitigate structural racism.[1]

In this paper, we raise a different kind of criticism, one that researchers and philosophers studying implicit bias must confront even if they can meet these formidable challenges. We will argue that the empirical research has not demonstrated, and without radical changes *could not* demonstrate, that the rapid associative behaviors it measures are anything more than straightforward expressions of subjects' *explicit* attitudes.

Fascination with implicit bias is predicated on the basic claim, offered by psychologists and accepted by philosophers, that the rapid associations and behavior that tests such as the Implicit Association Test (IAT) measure are *not* merely expressions of subjects' explicit attitudes. We will call this fundamental assumption, which drives the need to posit some other set of exotic attitudes to explain the data, the *Implicit Explanation*. The Implicit Explanation implies conclusions that are both surprising and morally significant, and that further fuel public and philosophical interest in implicit bias. Perhaps the most striking and troubling of these is the claim that even egalitarian people, who hold egalitarian explicit attitudes and beliefs, are often influenced by implicit mental processes that operate independently from, and are largely insensitive to, their explicit attitudes. Depending on how one interprets this conclusion, one may deduce from the empirical research that, disturbingly, no matter our moral convictions, the sub-personal processes of cognition that help us navigate the world will lead us to perform sexist and racist actions.[2]

The promise that empirical psychologists, armed with a new set of tools, would now be able to *measure* the operation of these sub-personal processes and ground this body of research in empirical fact has been another source of enthusiasm. We will argue here, however, that the data from the most prominent existing tests for implicit bias do not support the Implicit Explanation. Our argument will rely on the intuitive idea that some rapid mental associations, and the behaviors that such associations cause, express and reflect our explicit attitudes. For example, a teenager's immediate and reflexive lunge for a slice of pizza might express his love of pizza; a father's instantaneous grimace of pain when he sees his child skin her knee reflects his concern for her; a classical pianist's tendency to associate the concept [MUZAK] with [BAD] is an expression of her belief that Muzak is a grating abomination. These agents are open books: their rapid associations and behaviors tell us something about their explicit attitudes. If the rapid associations and behavior measured by existing implicit bias tests were also straightforward expressions of agents' *explicit* attitudes in this way, then the Implicit Explanation would be false. In this paper, we argue that implicit bias research has not undermined this banal Explicit Explanation of implicit bias test results.

To demonstrate that empirical research has failed to establish the Implicit Explanation, we will begin by showing that researchers have conducted implicit bias testing

[1] For example, see Oswald et al. (2013), Machery (2017, 2021), and Haslanger (2015) for representatives of each criticism, respectively. Prominent examples in popular press include Singal (2017) and Bartlett (2017). Brownstein, Madva and Gawronski (2020) respond to various criticisms within empirical psychology, the popular press, and philosophy. They respond to an online discussion of ours, but we believe that they misconstrue our argument. We will clarify our argument here.

[2] See for example, Jennifer Saul (2012), and Louise Antony's (2016) response.

in light of inadequate theories of, and therefore without adequate controls for, *explicit prejudice*. As a result, their data may merely reflect subjects' explicit prejudice that extant tests simply fail to measure. Next, we will offer suggestions for how researchers might improve their tests to attempt to address the criticisms we raise, while highlighting the underappreciated difficulties of interpreting the data that even the best possible empirical testing would yield. Finally, we will expand upon the philosophical significance of our criticism. We will show that in their eagerness to chart its philosophical and ethical implications, philosophers have not been sufficiently skeptical of the Implicit Explanation. As a result, they have encouraged a larger social shift in our understanding of how to relate to prejudiced agents that has serious moral costs. We conclude with the positive suggestion that philosophers interested in implicit bias shift their focus toward developing and defending views of explicit prejudice that could serve as the basis of successful controls in implicit attitude testing.

2 Measuring implicit bias

Before making the case that implicit bias research has failed to vindicate the Implicit Explanation, we note two caveats that will frame our discussion:

1. Most of the current work within social and empirical psychology, as well as much of the philosophical literature on implicit bias, relies on data obtained from several dominant experiments conducted and described in frequently cited papers. Most prominent is the "Implicit Association Test" (IAT), developed by Greenwald et al. (1998). Other dominant tests include the "Go/No-Go Association Task" ("GNAT"; described in Nosek & Banaji, 2001), the "Sorting Paired Features Task" ("SPF"; described in Bar-Anan et al., 2009), the "Brief Implicit Association Test" (BIAT; described in Sriram et al. 2009), the "Affect Misattribution Procedure" ("AMP"; described in Payne, 2009), and the "Affective Lexical Priming Score" ("ALPS"; described in Lebrecht et al., 2009). We also focus our discussion on these tests.
2. We use the term "implicit bias" to refer to bias that is at odds with a subject's explicit attitudes.[3] In pressing our criticism of implicit bias research, we treat "explicit attitudes" as attitudes that subjects could verbally report if prompted to do so. Common theoretical interpretations of "implicit attitudes" tend to assume that because subjects cannot verbally report their presence, or assess their influence on behavior, such attitudes must be "unconscious," "uncontrollable," or "arational."

[3] One might classify "implicit attitudes" in various ways, but these terminological differences will not affect our basic argument. One might, for example, hold that implicit attitudes are just those attitudes that subjects exhibit in implicit attitude testing. On that understanding of "implicit," our argument could be reformulated as follows: Empirical research on implicit bias has not shown that implicit biases, in this sense, are not straightforward expressions of subjects' *explicit* biases. Similarly, if one held that implicit attitudes are just those attitudes that *must* be measured by implicit attitude tests and could not, in principle, be verbally reported, then our argument could be restated as follows: Empirical research on implicit bias has not demonstrated the existence of biased implicit attitudes because it has failed to provide evidence that the attitudes subjects express in implicit bias testing could not, in principle, be verbally reported. In any case, our fundamental thesis is the same: Empirical research on implicit bias has not demonstrated that subjects' rapid associations and behavior are not straightforward expressions of their explicit attitudes.

To demonstrate the truth of the Implicit Explanation, one would need to provide good evidence that subjects' explicit attitudes did not underlie and explain their rapid associations and behavior. The most straightforward way to do so would be to show that even test subjects who harbored no explicit prejudice still exhibited biases in their rapid associations and actions. Indeed, this is what standard interpretations of implicit attitude testing claim that such testing does.[4] Once a test established that a subject did not harbor explicit biases, one could safely conclude that her biased associations and actions in test conditions were best explained by implicit mental processes. But, as we will now argue, the dominant implicit bias tests fail to demonstrate such a disparity and therefore fail to demonstrate the existence of implicit bias. The reason is straightforward: such testing has been conducted with inadequate measures of subjects' explicit prejudices.

The most common measures of explicit prejudice the dominant tests employ are the "Feeling Thermometer" and the "preference" survey. The first involves subjects' self-reports of "temperature" on a numerical scale. For example, when filling out a Feeling Thermometer for implicit racial prejudice, participants are asked, "How warm or cold do you feel toward Black people?" and "How warm or cold do you feel toward White people?" On preference tests for implicit racial bias, participants are asked which of a set of statements best describes them: "I strongly/moderately/slightly prefer White people (or White Americans) to Black people (or African Americans)" or, "I strongly/moderately/slightly prefer Black people to White people," or "I like White and Black people equally."[5]

The preference and Feeling Thermometer surveys are *the standard and dominant* explicit attitude tests typically cited by both empirical psychologists and philosophers who write about implicit bias. In fact, preference and temperature surveys are the *only* means the canonical versions of the SPF (Bar-Anan et al., 2009, p. 333) and the GNAT (Nosek & Banaji, 2001, p. 651) use to measure explicit prejudice.[6] Moreover, it is worth emphasizing that many studies draw exclusively on data obtained from the IAT as administered by the Project Implicit "virtual laboratory" (Rachlinski et al., 2009; Beaman et al., 2009, Sabin et al., 2009, for example). More than 20 million people have taken the IAT via the Project Implicit "Virtual Laboratory," where racial implicit attitude tests have no controls for explicit prejudice beyond standard Feeling Thermometer and preference measures.[7]

These extant tests of "explicit prejudice" are insufficient to detect, and so to control for, explicitly prejudiced attitudes. This is because a person may sincerely feel and report "warmth" toward members of the group he is prejudiced against. Consider, for

[4] See, for example, (Greenwald et al., 1998, p. 1475).

[5] These questions are taken directly from the IAT as administered by the Project Implicit website, but they are also included in the academic studies that describe the explicit measures used. See, for example, (Sabin et al., 2009).

[6] Some measures of implicit bias, such as the ALPS (Lebrecht, 2009), were not constructed using *any* explicit measures at all, presumably on the assumption that implicit attitudes can be measured directly and without controlling for explicit attitudes.

[7] In their 1998 paper introducing the IAT, Greenwald et al. employed a more sophisticated measure of explicit bias that we will discuss in the next section. This control was abandoned in many future uses of the IAT, including the version accessible via the Project Implicit Virtual Laboratory, which Greenwald co-founded.

example, a form of prejudice that involves holding explicit beliefs about a person's proper place in a social or natural hierarchy. Imagine, for example, a man who explicitly believes that women are goddesses who should be put on a pedestal, who should be pampered at home, but who lack the natural aptitude necessary to hold public office. Such a man could coherently, and sincerely, report feeling the same "temperature" toward both men and women, or even that he "feels warmer" toward women. Or imagine a person who explicitly believes stereotypes about white intellectual superiority and black physicality, who coherently and sincerely reports having no "preference for," or feeling any difference in "warmth" toward, members of one racial group over the other.

To see just how poor these measures of explicit prejudice are, consider Jones, who explicitly believes that each race has its place in the world, with whites occupying the top rung. While Jones believes that non-whites are by nature inferior, he also holds that God loves people of all races. Jones is aware of the content of his beliefs, how he came to have these beliefs (e.g. he knows he developed them during childhood), and which of his actions result from these beliefs (e.g. he knows that he is harsher and sterner with young Black boys than with young white boys because the former "need more discipline" to "learn their place"). We can imagine, too, that Jones's beliefs about the races play some role in shaping the patterns of emotions and desires that he feels. Moreover, we can imagine that Jones fills out the thermometer and preference questionnaires slowly and deliberately, and that he reasons to certain conclusions using basic rules of logical inference, exhibiting characteristics which indicate that his actions are the result of his explicit attitudes. When Jones is asked how warmly he feels about whites and Blacks, he reasons, "White Man was made in God's image; Black Man was made by God as a lesser being whom White men are spiritually and morally obligated to help. Ultimately, we are all God's creatures. 10/10 for both." We can imagine, too, that Jones's beliefs about the races play some role in shaping the patterns of emotions and desires that he feels.

Let us assume that Jones forms prejudiced conceptual associations that involve the concepts [BLACK] and [WHITE]. Because these prejudiced conceptual associations are totally consistent with his explicit attitudes, it would be surprising if Jones's rapid associations, and whatever actions they cause, were anything more than expressions of his explicitly racist beliefs. But given the inadequacy of the dominant tests' controls for explicit prejudice, Jones would be classified as a subject who holds "egalitarian beliefs" but holds "implicitly" prejudiced attitudes toward Blacks—a clearly absurd conclusion.

It should go without saying that one's explicit racism need not be as obvious as Jones's for the point to hold. Subjects whose prejudices are subtler than Jones's, but nevertheless still explicit, will easily evade the standard controls as well. To infer the Implicit Explanation from results on the dominant implicit bias tests would not be justified.

3 Meeting the theoretical challenge?

To develop a test that could serve as a control for subjects' explicit biases, one would need to know what kinds of explicit attitudes could potentially underlie and explain subjects' biased associations and behaviors. This involves commitment to theoretical claims about what explicit prejudice consists in. Use of the Feeling Thermometer, for example, implies that feelings of relative "coldness," and no other explicit attitudes, could explain biased rapid associations and behavior in test conditions. We have argued that the Feeling Thermometer is too crude to rule out the Explicit Explanation of such behavior, but it is not obvious what an adequate supplement to that rudimentary model of prejudice might be. In this section, we will present several conceptual and practical difficulties researchers would have to overcome to construct and use an acceptable alternative. We do not claim that building a satisfactory model would be impossible, but, as we will show, the current research is not close to meeting the challenges we will outline.

One difficulty is that a successful theory of prejudice must vary depending on the form of prejudice one is controlling for and the characteristic operations of that prejudice in the local context one is studying. Explicit racism toward Blacks in the southern United States, for example, may characteristically consist in a set of beliefs that differs from the beliefs held by explicitly prejudiced subjects in northern states. And for any general form of prejudice (racism, sexism), there will be competing conceptions of what that form of prejudice consists in. Is racist prejudice, for example, fundamentally a matter of holding different attitudes toward different racial groups (as the Feeling Thermometer seems to assume), a matter of holding attitudes that evince acceptance or endorsement of racial domination, or something else altogether? Does sexist prejudice consist in holding different attitudes toward men and women, holding attitudes that implicate one in patterns of objectification or sex-based domination, or in something else altogether?

Moreover, and to state the obvious, the mechanisms and psychological manifestations of racist and sexist prejudice in each context are complex and evolving. For example: Would recording the negative stereotypes participants held about members of a marginalized group be sufficient to supplement the Feeling Thermometer, or would one have to ask about "positive" stereotypes about members of that group as well?[8] A good answer would draw on a theory of prejudice and its operations, and one could press similar questions about other attitudes and beliefs. Should researchers ask subjects about their positive beliefs about members of the dominant group? Their political ideals? Their appreciation of historical injustices? Their hopes and fears? Their senses of humor? Answering these questions, and determining the extent to which rapid associations and behaviors may be straightforward expressions of explicitly prejudiced attitudes, requires an intellectually serious theory of sexist and racist moral psychology and a plausible conception of the functioning of sexism and racism in the society under scrutiny.

[8] Of course, this assumes that one knows which stereotypes to ask about. Such knowledge would require a good theory of prejudice and insight into the sociopolitical dynamics of the local context.

Though they are rarely used to control for explicit attitudes in implicit bias research, there are psychological tests that reflect more theoretically sophisticated views of racism and sexism and thus seem to acknowledge some of these difficulties.[9] These include John McConahay's (1981) Modern Racism Scale, the Symbolic Racism Scale (Henry & Sears, 2002), the Diversity Scale (Wittenbrick et al. 1997), and the Discrimination Scale (Payne et al., 2009). Similarly, the Ambivalent Sexism Inventory (Glick & Fiske, 1996) clearly reflects a conception of sexism that goes beyond "temperature" and preference.

To be clear, this is not to say that these measures of explicit bias are "better" in the sense that they more accurately *predict* discriminatory behavior than temperature or preference scales. Rather, they are *conceptual* improvements.[10] These more complex measures of explicit racism and sexism would make for better explicit bias controls because they strive to identify and measure ways in which a subject could harbor forms of explicit prejudice that would evade cruder surveys but that could still underlie and explain the subject's rapid associations and behavior. For example, the Ambivalent Sexism Inventory presumes, rightly, in our view, that sexism is not merely a matter of "temperature" and preference, but, rather, that it can be constituted by more complex attitudes, many of which could be "positive," "warm," or, as the developers of the test write, "benevolent" ("A good woman should be set on a pedestal"; "Women have a quality of purity few men possess"). In selecting from among possible controls, researchers are, in effect, taking a position on the question of which explicit attitudes could underlie and explain subjects' rapid associations. Again, this implies a commitment to a substantive theory of racism or sexism.

To be fair, some implicit bias researchers have used the more theoretically sophisticated measures we mentioned above.[11] But we will argue that even implicit bias tests that do employ these more robust controls fail to vindicate the Implicit Explanation.

To fully evaluate whether even the best existing tests are reliable indicators of explicit prejudice, one would first need to successfully present a justifiable and plausible theory of explicit prejudice. Then, one would need to show that some set of existing tests could, with a reasonable degree of accuracy, identify those respondents with views marking them as explicitly prejudiced according to the theory. Implicit bias researchers have not done this.

[9] We include the full content of these measures in the "Appendix". Where full content was unavailable, we've included the "Representative Items" offered by the original authors.

[10] Brownstein et al. (2020), responding to a blog post based on an unpublished version of this manuscript, characterize our skepticism as grounded in the worry that implicit measures are "poor predictors" of prejudiced behavior, and that explicit measures may serve as better predictors. This is a misunderstanding of our point. We are not arguing that implicit bias tests fail to predict behavior. Rather, our point is that empirical research has not shown that the behavior implicit bias tests measure, predict, or explain is at odds with subjects' explicit attitudes. Our argument does not rest on a doubt that tests such as the IAT measure subjects' current rapid associations, or that, as Brownstein et al. put it, implicit measures such as the IAT "reflect what is going on in a person's mind in a given moment, which is shaped by complex interactions of person-related and situation-related factors (296)." Again, what we doubt is that what is measured by the IAT is best understood as discrepant with a subject's explicit attitudes. The IAT may indeed reflect what is going on in a person's mind at a given moment, but what is going on in her mind at that moment may be an operation of her explicit prejudice.

[11] See Greenwald et al. (1998) and Payne et al. (2009).

Even if implicit bias researchers did show that the most sophisticated existing surveys (or some combination of them) reflected the best theories of the biases they were studying, they would still have to overcome serious difficulties of interpretation in order to use them to control for explicit bias. Crucially, they would have to attend to the distinction between *measuring* explicit prejudice and *controlling* for it on an implicit bias test. Controlling for explicit prejudice requires ruling out the Explicit Explanation for implicit bias test results, not determining "how racist" or "how sexist" respondents are. All of the most sophisticated tests aim to measure subjects' *degrees* of prejudice, but the Implicit Explanation is a claim about the *foundations* of certain associative and behavioral patterns. This means that while one might be justified in concluding, on the basis of a low overall score on the Modern Racism Scale, for example, that a subject harbored only a small degree of explicit racial bias, one would not be justified in concluding that the handful of prejudiced explicit attitudes the subject *did* harbor did not best explain his biased associations and behavior in test conditions.

The way psychologists tend to analyze the data they obtain from these tests virtually guarantees that they will not be able to rule out the Explicit Explanation. For example, in Greenwald et al. (1998), researchers scored subjects' responses to questions about their racial attitudes on a scale of 1 to 5, with lower scores indicating less explicit prejudice. Thus, even a person receiving a very "low" overall score might, for example, *strongly agree* that "Blacks should not push themselves where they are not wanted."[12] If such a person then registered biased rapid associations, would his answers to the explicit bias control rule out the Explicit Explanation? We think not. Despite his low overall score, it is plausible that this single prejudiced explicit attitude could best explain his rapid associative patterns. Or suppose a subject who took the Discrimination Scale (Payne et al., 2009) reported that he did not associate Black people with any of the listed stereotypes except the stereotype that they are violent, or the stereotype that they are not "intelligent at school." Endorsing one or two stereotypes of this kind would not prevent him from registering a low overall score on the scale, but it could easily explain a host of biased associative patterns and actions.

Consideration of this issue raises a more general question of interpretation: What do subjects' answers to these more sophisticated questions *mean*? What is the ethical significance, for example, of someone "agreeing somewhat" or "disagreeing slightly" with statements such as "Once a man commits [to a woman], she puts him on a tight leash," (Glick & Fiske, 1996) or reporting that he has "felt sympathy for Blacks about half the time?" (Payne et al., 2009) What does it say about a person if he believes that "There is a real danger that too much emphasis on cultural diversity will tear the United States apart" (Wittenbreck et al., 1997), that "Black leaders have been moving at about the right speed" (Henry & Sears, 2002), or that "Women have a superior moral sensibility" (Glick & Fiske, 1996)?

In highlighting these possible questionnaire responses, we do not mean to suggest that the questions they answer are the wrong kinds of questions to ask. On the contrary, they may yield just the kind of data one would need to assess a person's explicit racist or sexist attitudes. But the very complexity and nuance that distinguishes tests that ask such questions from their simpler counterparts make them difficult to analyze. A

[12] From "The Modern Racism Scale"; this question is a sample from Greenwald (1998).

successful interpretation of a respondent's answers to these loaded questions would require an excellent understanding not only of how prejudice functions in the context one is studying, but also of how the subject's responses relate to his own psychology and background. Someone like Jones, for example, might feel sympathy for Black people all the time. He sees them as lowly and pitiable, after all. To score this response as evidence of *anti*-racism, or egalitarianism, when it is in fact just the opposite, would be a mistake. And a participant who agreed that women tended to keep men on "tight leashes" might be offering a report of her life experience, endorsing a misogynistic stereotype, or both. All of the most sophisticated tests for explicit racist and sexist prejudice are subject to these worries.

At this point, one might ask whether, in the absence of effective controls for explicit prejudice, and in light of these interpretive difficulties, there is other evidence in favor of the Implicit Explanation, or if there are alternative measures researchers could use to avoid some of the challenges we've highlighted here. For example, it may seem significant that, anecdotally, many people who take implicit association tests are surprised by their results, and one may wonder if such surprise is itself evidence for the Implicit Explanation. A genuine egalitarian, after all, would have reason to be taken aback upon learning that he formed prejudiced conceptual associations under test conditions.

Researchers have not attempted to use surprise as a control for explicit attitudes, but we might still ask whether such an approach could yield data that supported the Implicit Explanation.[13] We doubt, however, that such a strategy would succeed, given that there are many reasons why one might be surprised by implicit bias test results that do not imply a discrepancy between explicit attitudes and rapid associations. We will offer just two here.

1. Subjects who "fail" implicit bias tests such as the IAT do so because they fail to perform a task the tests instruct them to perform, such as quickly sorting photographs into categories. Those who are surprised at their results may simply be surprised at their failure to complete the tasks successfully, which is compatible with being surprised that they were unable to prevent their *explicit* attitudes from influencing their rapid associative behavior. We might see such surprise outside of test conditions, as well. A person might be surprised, for example, that he could not suppress tears of joy at his child's wedding, even though he tried very hard to keep his composure. But his surprise would not be evidence that his tears were anything but an expression of his explicit love and happiness.
2. A person may possess an inadequate conception of what "egalitarianism" requires and believe, wrongly, that he holds egalitarian explicit attitudes when, in fact, he does not. This misconception could explain a test subject's sense of surprise at his results without appealing to a mismatch between his rapid associations and explicit attitudes. Imagine a man who believes that women have the right to work outside the home and to control their own bodies, but that as a matter of "scientific fact," men make for better leaders. Such a man might be surprised by his propensity to associate [WOMAN] and [HOME] on an implicit attitude test, but this test result is not at odds with his explicit attitudes.

[13] We are grateful to an anonymous reviewer for raising this suggestion.

We are not claiming that these *are* in fact the best explanations of subjects' surprise at their test results. Rather, these alternate explanations aim to undermine the inference from surprise to the Implicit Explanation. Measures of surprise cannot take the place of robust controls for prejudiced explicit attitudes that reflect sophisticated conceptions of explicit prejudice and its operation in particular contexts.

To be clear: our position is not that it is in principle impossible to effectively control for explicit bias.[14] Talented researchers may be able to overcome the obstacles we have described. But we do believe that the challenges are formidable. Meeting them, and vindicating the Implicit Explanation, would require conceptual insight and ingenious test construction that the current empirical research has not achieved.

4 Implications and clarifications

We will now discuss the implications of this failure, both for philosophy and for society at large. The two are, of course, related, and it is easy to see why implicit bias research has excited both theorists and laypersons. Implicit bias scholarship purports to use empirical methods to demonstrate a surprising moral-psychological claim with enormous ethical and social implications: Even good people with good explicit attitudes ("committed egalitarians") are subject to rapid associations and behavior that cause them to behave in prejudiced ways. If true, this would indeed be a disturbing conclusion—one that would call for us to rethink the ways in which we understand the operations of prejudice and our own psychology. We have argued that this claim, which has driven both academic scholarship and public discourse, has not been established.

At this point, one might protest that at least some of the recent criticism of implicit bias research, including the critique that we are raising here, is misplaced. One might suggest that we are reacting to a *misrepresentation* of the significance of implicit bias research by media members unfamiliar with the underlying science, by opportunistic public figures, or by institutions eager to employ "implicit bias training" in lieu of addressing deeper issues. We do not deny that such figures have sometimes exaggerated and misinterpreted the significance of the empirical research. But our criticism does not target the embellishments and "hype" in public discourse that psychologists and

[14] Thus, though we are sympathetic with Eduoard Machery's (2017) suggestion that implicit bias research has attempted to construct theories "on quicksand," we nonetheless think that our particular criticism may be addressable with the right kind of philosophical intervention. Machery (2016) argues that the distinction between "implicit" and "explicit" attitudes is meaningless, because bias should be understood as a character trait. It thus makes no sense to draw a distinction between implicit and explicit bias. It is consistent with our argument, however, that it may make sense to think of prejudice as a trait that could, in principle, be partly constituted by either explicit or implicit attitudes. And this leaves open the possibility that empirical research may be attempting to show the surprising and morally significant thesis that people who have egalitarian explicit attitudes may nonetheless possess prejudiced implicit attitudes. Thus, unlike Machery (2016), we are not arguing that empirical psychologists have been attempting to demonstrate a thesis that relies on a conceptual confusion. Rather, we are attempting to show that they have not demonstrated the truth of that thesis. Our thanks to a reviewer for suggesting this clarification. For a recent presentation of methodological criticisms of implicit bias research that are distinct from the critique we press here, see Machery (2021).

others familiar the empirical literature would reject.[15] Rather, we are casting doubt on the *basic claim* that the biases the empirical tests reveal are implicit.

In doing so, we are also responding to the widespread uncritical acceptance of the Implicit Explanation among *philosophers.* Philosophical articles on implicit bias tend to remind their audiences of the significance of the topic by echoing the core claim of the empirical research, that it has provided us with good evidence of the Implicit Explanation, before going on to theorize about what could explain the discrepancy between explicit and implicit attitudes, or about the moral significance of this discrepancy. Here are some prominent examples:

1. "Research on implicit bias demonstrates that individuals can act in discriminatory ways even in the absence of explicitly prejudiced motivations" (Brownstein and Madva, 2018, p. 1).
2. "Evidence has been building that implicit attitudes are at least moderately good at predicting real-world behavior, independent of the effects of people's explicit (verbally reported) attitudes" (Carruthers, 2018, p. 1).
3. "What explains the apparent disparity between self-reported attitudes and behaviour?… While self-presentation effects undoubtedly play a role, it is very likely that what psychologists call *implicit* attitudes explain some of the disparity between reported attitudes and behaviour" (Levy, 2017, p. 535).
4. "There is abundant evidence that most people, often in spite of their conscious beliefs, values, and attitudes, have implicit bias. 'Implicit bias' is a term of art referring to evaluations of social groups that are largely outside of conscious awareness or control" (Brownstein & Saul, 2016, pp. 1–2).

If our argument is sound, then even this *basic framing* of the topic, which begins with and relies on the conclusion that implicit bias tests have demonstrated a discrepancy between rapid associations and explicit attitudes, is misleading.

Once accepted, the Implicit Explanation gives rise to at least three kinds of philosophical questions that philosophers have been eager to take up. The first set of questions is metaphysical: What are these implicit biases that help to shape our behavior? Are they best understood as emotions, beliefs, "aliefs," "behavioral clusters," or something else altogether? And given what they are, are they unconscious, arational, or uncontrollable? One's answers to these metaphysical inquiries lead to the second and third topics, which are both ethical: First, are we morally responsible, and therefore potentially blameworthy, for our biased rapid associations and behavior? Second, how should we go about altering our pernicious associative patterns, given our moral and political hopes of living in a more just society? We will consider these three areas, metaphysics, responsibility, and intervention, in turn.

Metaphysics: Credulous acceptance of the Implicit Explanation by the public and philosophers has been mutually reinforcing and has helped shape public thought about human psychology and the operations of individual prejudice. More specifically, it has

[15] As an example of the "hyping" of implicit bias research, Brownstein et al. (2020) offer Nicholas Kristof's claim that, "It's sobering to discover that whatever you believe intellectually, you are biased about race, gender, age, or disability." They suggest that this conclusion is not informed by the actual science, which suggests that "explicit beliefs about social concepts are, in fact, strong moderators of implicit attitudes about those concepts" (Brownstein et al., 2020, p. 298, fn. 15).

helped establish and strengthen a conception of human psychology and prejudice that understands rapid associative behavior as a mental phenomenon divorced from, and potentially unresponsive to, the explicit attitudes that one might hope would form the basis of our self-understandings and our everyday interactions with other people. This conception depends on a particular kind of metaphysical interpretation of what underlies and explains rapid associative behavior, one that understands it as the product of a cognitive architecture that operates automatically, arationally, or outside of our conscious control.[16]

Moral responsibility: This understanding of the relationship between rapid associative behavior and human psychology has ethical implications. Notably, it is relevant to inquiry about our moral responsibility for rapid associative behavior and its consequences. Insofar as one's metaphysical interpretation of rapid associative behavior distinguishes it from attitudes and beliefs that we rationally endorse or control, one may be moved to take up what P.F. Strawson called the "objective," rather than "participant,"[17] stance toward "implicitly" biased agents (including oneself), emphasizing treatment and management as opposed to reasoning, conversation, empathetic engagement, and the feeling and expression of "reactive attitudes."

We realize that this characterization of the pressure to move away from the interpersonal ideals of the participant stance is abstract. There are, of course, competing views of responsibility for "implicit" bias in the philosophical literature, each of which has its own subtleties. To mention a few representative examples: Natalia Washington and Daniel Kelly (2016) argue that individual moral responsibility for implicit bias turns on whether or not knowledge of the empirical science surrounding implicit bias is available in that person's community. If not, then the person should not be held fully accountable and blameworthy for his behavior. Maureen Sie and Nicole van Voorst Vader-Bours (2016) urge a re-orientation from individual responsibility toward collective responsibility for implicit bias. And Robin Zheng (2016) argues that we may hold "implicitly" biased agents accountable for their behavior in some sense (for example, by demanding that they compensate victims, make efforts to change their behavior in the future, or make amends), but that blaming them "would be like blaming a person for a behavior that they acquired as the result of some trauma, which gets triggered under certain circumstances; while such a disposition is something to be managed by her and others, it is not something for which she deserves blame or deep moral criticism" (79).[18]

Our aim here is not to evaluate the comparative merits of these metaphysical and ethical positions. Rather, we are focusing on what they share in common to make two general points. First, that one's metaphysical interpretation of rapid associative

[16] For example, consider interpretations such as Madva and Brownstein's (2018) proposal that implicit attitudes are clusters of semantic-affective associations that can be trained to change over time, but tend to be unresponsive to the semantic content of our other mental states, Gendler's proposal that, like phobias, implicit attitudes are explained by "aliefs" (Gendler, 2008, 2011), Levy's proposal that they are "patchy endorsements" that resist rational correction, and the view that they are the result of "system 1" processes on a dual system theory of mind (Gawronski & Bodenhausen, 2006).

[17] Strawson (1962).

[18] As Zheng then elaborates in a footnote, "Here, with respect to this particular trait, we adopt the Strawsonian 'objective' attitude, the attitude we take towards non-human animals, young children, and beings that are not fully moral agents" (79, fn. 30).

behavior can make a concrete difference to our everyday interactions and relationships. Second, that acceptance of the Implicit Explanation, combined with standard views about moral responsibility, tends to encourage at least some shift away from the participant stance and its standard modes of interaction.

If the empirical research provided strong evidence for the Implicit Explanation, then one would need to carefully consider the merits of these views of responsibility for rapid associative behavior. If, on the other hand, the Explicit Explanation turned out to be correct (or largely correct), then we could reject all of them, and therefore avoid the serious moral and social costs of a shift toward the objective stance. If the Explicit Explanation is true, then to tell victims that transgressors' behavior cannot be attributed to them, or that they are not proper targets of "deep" criticism and blame, would not only be misleading; it would deprive victims of the opportunity to confront and engage with offenders in a way that fully reflects their shared humanity. And to claim that offenders' prejudiced associations are merely regrettable products of their cognitive architecture and their surroundings, rather than expressions of their rational agency, would not only be a distortion but an insult to the transgressors themselves.

Intervention: A similar point applies to efforts to alter prejudiced behavior. Acceptance of the Implicit Explanation may lead one to suspect that rational changes to rapid associative behavior are difficult or even impossible. And indeed, some recent work on implicit bias in empirical psychology has focused on how to mitigate or intervene on implicit bias using non-rational methods that rely on re-conditioning agents' associations or encouraging subjects to exercise more control over their actions in order to align them with their egalitarian beliefs.[19]

But while one might be able to "recondition" one's association of dark skin with the concept [DANGEROUS], this change would have nothing to do with one's *understanding, perception,* or *outlook toward* people with dark skin. Similarly, it would be one thing to eliminate the stereotype that women are bad at math through retraining evaluative-semantic associations between the phrase "Women are good at" and "math." It would be another thing to eliminate one's belief that women are naturally bad at math by coming to a *realization* that women can be good at math, and to experience confidence in their mathematical abilities as a result.

To be clear, our concern is not primarily prudential but *ethical*, motivated by a moral commitment that is deeply embedded within the broadly liberal outlook that inspires standard objections to sexism and racism in the first place. The point is not that "rational" interventions would be more effective than non-rational interventions. Rather, our concern is that these non-rational methods, when aimed at agents whose rapid associations do in fact reflect their explicit attitudes, are inconsistent with respect. They merely manipulate a person's behavior, rather than encourage rational changes in how he views and responds to others.

[19] For example, Kawakami et al., (2005, 2007) have tested the effects of "counterstereotype training," which involves having subjects respond "Yes" to images of Black people or women, or having subjects nod (Wennekers, 2012) or pull a joystick toward themselves when prompted with Black or Arab-Muslim faces and push it away from themselves in response to White faces. Forbes and Schmader (2010) tested the effects of training subjects to associate the phrase "women are good at" with math terms, as well as interventions that aim to reduce the influence of one's implicit biases on one's actions without intervening on one's psychology (Beauclac and Kenyon, 2014). For a defense of using these forms of de-biasing techniques outside the laboratory, see Madva (2017).

If the Implicit Explanation does turn out to be true, and re-conditioning is needed, this would be a disturbing fact, and one we should not take lightly. The conclusion that such means are necessary should be understood not as an opportunity to toast the dawning of an exciting new era of science-based approaches to moral improvement, but rather as a sobering concession that we cannot confront a new, pernicious form of racism and sexism without abandoning our traditional understanding of persons as responsible agents who could come to understand one another as moral equals. This should be seen for the pessimistic conclusion it is, regardless of the cheery progressive attitude that can sometimes accompany it. Our argument, that psychologists and philosophers in this debate have not shown that the biases in question are indeed implicit, gives us reason to eschew, or at least delay, such pessimism.

5 Conclusion

It is important that philosophers, especially moral psychologists, engage with empirical observation, and not only because their work should be informed by real life, and what real human beings are like. Philosophical reflection and conceptual refinement allow us to better interpret experimental data. In the case of implicit bias, philosophers have been eager to interpret empirical results, but they have largely focused their attention on answering the three kinds of questions we outlined above. In the process, they have produced a vast body of technical literature in a short period of time, offering critical and competing views of what implicit attitudes are and the implications that follow from these metaphysical conclusions. We have not compared the relative merits of these positions because our contention is that the philosophical drive to interpret needed to express itself at an earlier stage.[20]

Thus, though we are critical of the extant work on implicit bias, our thesis is not entirely negative. We believe that our understanding of the current research's conceptual shortcomings suggests a way in which both empirical psychologists and philosophers working on implicit bias might productively change course. Psychologists should devote attention toward finding better ways to measure (and not solely predict) explicit prejudice, rather than focusing primarily on refining their methods of tracking implicit associations and behavior. In doing so, they should bear in mind the points we raised in section III, where we discussed the difficulties of developing adequate controls for explicit prejudice and stressed that doing so would require both a sophisticated theory of the way that prejudice operates in the context under study and a sense of the agent's psychology and history. Developing effective controls for explicit prejudice will involve drawing on *normative* moral-psychological theories—theories

[20] Even philosophers who argue for metaphysical conceptions of implicit prejudice that challenge common conceptions of implicit bias assume that the empirical research has at least demonstrated a discrepancy between subjects' implicit and explicit attitudes. Consider, for example, Carruthers' (2018) proposal that the same mental structures can underlie and explain *both* "explicit" prejudice (what a subject would report about herself when prompted) and "implicit" associations and behavior. Carruthers may be right that we need not develop a special ontology to explain implicit bias test results. But even he grants that the empirical research has shown a discrepancy between what subjects would be willing to report if asked and their rapid associations. It is this basic assumption that we are critical of. Our thanks to an anonymous reviewer for encouraging us to distinguish our argument from Carruthers'.

that help us identify morally good and bad states of mind. Ethicists and normative moral psychologists must be seated at the table alongside the empirical psychologists and philosophers of mind who take on the challenge.[21,22]

Appendix

Where the full content was not made available by authors, we include the "Representative Items" listed.

Glick, Peter and Susan T. Fiske. "The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism." *Journal of Personality and Social Psychology*, 70(3), (1996).

[21] To be clear, we do not think that moral philosophers should simply ignore the findings of empirical psychologists, either. For an elaboration on the relationship between normative moral psychology and empirical psychology, see Wolf (2007).

[22] We are grateful for feedback we received from Susan Wolf, Douglas MacLean, Ram Neta, Robert Smithson, Charles Siewart, Calvin Lai, Elizabeth Reis, Matthew Dennis, Pamela Reis, Chris Hakkenberg, and the anonymous referees who reviewed the manuscript. We have also benefited from discussion with audiences at the Rocky Mountain Ethics Congress and at a meeting of the American Philosophical Association, where Elís Miller Larson commented on the paper. We would also like to thank Brad Cokelet and David Shoemaker for hosting an online discussion of the paper on the PEA Soup blog.

Scale item	Study 1	2	3	4	5
Hostile Sexism					
Women exaggerate problems at work	.71	.70	.71	.80	.73
Women are too easily offended	.76	.81	.66	.69	.66
Most women interpret innocent remarks as sexist	.74	.69	.61	.55	.70
When women lose fairly, they claim discrimination[b]	.74	.49	.31	.77	.66
Women seek special favors under guise of equality	.68	.74	.70	.59	.71
Feminists are making reasonable demands[a]	.75	.60	.50	.49	.42
Feminists not seeking more power than men[a]	.73	.50	.47	.56	.64
Women seek power by gaining control over men	.67	.64	.72	.70	.69
Few women tease men sexually[a]	.60	.51	.37	.25	.46
Once a man commits, she puts him on a tight leash	.73	.65	.65	.81	.77
Women fail to appreciate all men do for them	.68	.69	.66	.64	.58
Benevolent Sexism					
Protective Paternalism					
A good woman should be set on a pedestal	.68	.58	.66	.58	.62
Women should be cherished and protected by men[b]	.69	.43	.28	.66	.49
Men should sacrifice to provide for women	.69	.54	.73	.67	.69
In a disaster, women need not be rescued first[a]	.62	.48	.35	.33	.47
Complementary Gender Differentiation					
Women have a superior moral sensibility	.69	.74	.75	.77	.56
Women have a quality of purity few men possess	.82	.80	.82	.78	.61
Women have a more refined sense of culture, taste	.72	.69	.71	.67	.71
Heterosexual Intimacy					
Every man ought to have a woman he adores	.67	.57	.69	.64	.55
Men are complete without women[a]	.69	.70	.51	.63	.55
Despite accomplishment, men are incomplete without women	.79	.75	.84	.66	.71
People are often happy without heterosexual romance[a]	.67	.50	.37	.36	.44
N	811	171	937	144	112

Note. ASI = Ambivalent Sexism Inventory.
[a] Indicates items reverse-worded (and reverse-scored) for Studies 2–6 and on the final scale.
[b] Indicates items for which reversed wording (and reversed scoring) was used in Studies 2 and 3 but which were returned to their original wording for the final version of the scale and for Studies 4–6.

Greenwald, A., D. McGhee, and J. Schwartz. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test". *Journal of Personality and Social Psychology* 74 (1998): 1464–80.

Appendix B

Sample Items From Explicit Measures

Modern Racism Scale

Discrimination against Blacks is no longer a problem in the United States.

Over the past few years, the government and news media have shown more respect for Blacks than they deserve.

Diversity Scale

There is a real danger that too much emphasis on cultural diversity will tear the United States apart.

The establishment and maintenance of all-Black groups and coalitions prevents successful racial integration.

Discrimination Scale

Members of ethnic minorities have a tendency to blame Whites too much for problems that are of their own doing.

In the U.S. people are no longer judged by their skin color.

Note. All responses were scored from 1 to 5, with lower scores recoded to indicate less anti-Black prejudice.

Received January 2, 1997
Revision received September 11, 1997
Accepted September 15, 1997 ■

Henry, P.J. and David O. Sears, "The Symbolic Racism 2000 Scale," *Political Psychology*, 23:2 (2002).

APPENDIX: The Symbolic Racism 2000 Scale

1. It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites. (1, strongly agree; 2, somewhat agree; 3, somewhat disagree; 4, strongly disagree)
2. Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same. (1, strongly agree; 2, somewhat agree; 3, somewhat disagree; 4, strongly disagree)
7. Some say that black leaders have been trying to push too fast. Others feel that they haven't pushed fast enough. What do you think? (1, trying to push too fast; 2, going too slowly; 3, moving at about the right speed)
9. How much of the racial tension that exists in the United States today do you think blacks are responsible for creating? (1, all of it; 2, most; 3, some; 4, not much at all)
11. How much discrimination against blacks do you feel there is in the United States today, limiting their chances to get ahead? (1, a lot; 2, some; 3, just a little; 4, none at all)
12. Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class. (1, strongly agree; 2, somewhat agree; 3, somewhat disagree; 4, strongly disagree)
15. Over the past few years, blacks have gotten less than they deserve. (1, strongly agree; 2, somewhat agree; 3, somewhat disagree; 4, strongly disagree)
16. Over the past few years, blacks have gotten more economically than they deserve. (1, strongly agree; 2, somewhat agree; 3, somewhat disagree; 4, strongly disagree)

McConohay, John, Betty B. Hardee and Valerie Batts, "Has Racism Declined in America? It Depends on Who is Asking and What is Asked," *The Journal of Conflict Resolution* 25(4), 1981.

The "Modern Racism" Scale:

> It is easy to understand the anger of black people in America. (disagree)
>
> Blacks have more influence upon school desegregation plans than they ought to have. (agree)
>
> The streets are not safe these days without a policeman around. (agree)
>
> Blacks are getting too demanding in their push for equal rights. (agree)
>
> Over the past few years blacks have gotten more economically than they deserve. (agree)
>
> Over the past few years the government and news media have shown more respect to blacks than they deserve. (agree)

Payne, B. (2009). "Attitude Misattribution: Implications for Attitude Measurement and the Implicit-Explicit Relationship." In A. Black and W. Prokasy (Eds.) R. Petty, R. Fazio, and P. Brinol (Eds.), *Attitudes: Insights from the new wave of implicit measures*. Hillsdale, NJ: Erlbaum.

Sample 1 (ANES Panel Study)

Feelings: Do you feel warm, cold, or neither warm nor cold toward blacks? [Response options: Warm, cold, neither warm nor cold].

> If warm: Do you feel a little warm, moderately warm, or extremely warm toward blacks?
> If cold: Do you feel a little cold, moderately cold, or extremely cold toward blacks?

Sympathy: How often have you felt sympathy for blacks? [Always, most of the time, about half the time, once in a while, or never].

Admiration: How often have you felt admiration for blacks? [Always, most of the time, about half the time, once in a while, or never].

Influence: Would you say that blacks have too much influence in American politics, just about the right amount of influence in American politics, or too little influence in American politics? [Response options: Too much influence, Just about the right amount of influence, Too little influence].

Sample 2 (ANES Time Series)

Feelings: Do you feel warm, cold, or neither warm nor cold toward blacks? [Response options: Warm, cold, neither warm not cold].

> If warm: Do you feel a little warm, moderately warm, or extremely warm toward blacks?

If cold: Do you feel a little cold, moderately cold, or extremely cold toward blacks?

Sympathy: How often have you felt sympathy for blacks? [Always, most of the time, about half the time, once in a while, or never].

Admiration: How often have you felt admiration for blacks? [Always, most of the time, about half the time, once in a while, or never].

Stereotypes:

Where would you rate BLACKS on this scale? [1 = Hardworking; 7 = Lazy].

Where would you rate BLACKS on this scale? [1 = Intelligent; 7 = Unintelligent].

Symbolic Racism:

1. Irish, Italians, Jewish, and other minorities overcame prejudice and worked their way up, blacks should do the same without special favors" [Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree].
2. Generations of slavery have created conditions that make it difficult for blacks to work their way out of the lower class" [Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree].
3. It's really a matter of some people just not trying hard enough; if blacks would only try harder they could be just as well of as whites. [Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree].
4. Over the past few years, blacks have gotten less than they deserve." [Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree]

Sample 3 (Associated Press/Yahoo! News/Stanford University study)

Liking: "How much do you like or dislike each of the following groups? Whites … Blacks …" [Response options: dislike a great deal, dislike a moderate amount, dislike a little, Neither like nor dislike, like a little amount, like a moderate amount, like a great deal].

Admiration: "How often have you felt admiration for blacks?" [Extremely often, Very often, Moderately often, Rarely, Never].

Sympathy: "How often have you felt sympathy for blacks?" [Extremely often, Very often, Moderately often, Rarely, Never].

Stereotypes: Respondents were asked "How well does each of these words describe most blacks?" and were shown a list of 14 adjectives (*Friendly, Determined to succeed, Law abiding, Hard-working, Intelligent at school, Smart at everyday things, Good neighbors*, *Dependable, Keep up their property, Violent, Boastful, Complaining, Lazy, Irresponsible)*. [Extremely well, very well, moderately well, slightly well, not well at all].

Wittenbreck B., CM Judd, and B Park. "Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures." *Journal of Personality and Social Psychology*. 72(2):1997.

Table 1
Explicit Prejudice Measures: Representative Items and Coefficient Alpha Reliabilities

Representative item	Alpha
Modern Racism Scale (McConahay et al., 1981)	.802
"Blacks are getting too demanding in their push for equal rights."	
"Over the past few years, the government and news media have shown more respect to Blacks than they deserve."	
Pro-Black scale (I. Katz & Hass, 1988)	.771
"Black people do not have the same employment opportunities that Whites do."	
"Too many Black people still lose out on jobs and promotions because of their skin color."	
Anti-Black scale (I. Katz & Hass, 1988)	.859
"On the whole, Black people don't stress education and training."	
"The root cause of most of the social and economic ills of Blacks is the weakness and instability of the Black family."	
Diversity Scale	.672
"There is a real danger that too much emphasis on cultural diversity will tear the United States apart."	
"The desire of many ethnic minorities to maintain their cultural traditions impedes the achievement of racial equality."	
Discrimination Scale	
"Blacks are ultimately responsible for the state of race relations in this country."	
"More and more, Blacks use accusations of racism for their own advantage."	.885

References

Antony, L. (2016). Bias: Friend or Foe? Reflections on Saulish Skepticism. In J. Saul & M. Brownstein (Eds.), *Implicit bias and philosophy, volume I: Metaphysics and epistemology.* Oxford University Press.

Bar-Anan, Y., Nosek, B., & Vionello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology, 65*(5), 329–343.

Bartlett, T. (2017). Can we really measure implicit bias? Maybe not. *The Chronicle of Higher Education.* http://www.chronicle.com/article/Can-We-Really-Measure-Implicit/238807

Beaman, L., Chattopadhyay, R., Duflo, E., Rande, R., & Topalova, P. (2009). Powerful women: Does exposure reduce bias? *The Quarterly Journal of Economics, 124*, 1497–1540.

Beauclac, G., & Kenyon, T. (2014). Critical thinking education and debiasing. *Informal Logic, 34*(4), 341–363.

Brownstein, M., Madva, A., & Gawronski, B. (2020). "Understanding implicit bias: Putting criticism into perspective. *Pacific Philosophy Quarterly, 10*, 276–307.

Brownstein, M., & Saul, J. (2016). Introduction. In J. Saul & M. Brownstein (Eds.), *Implicit bias and philosophy, volume I: Metaphysics and epistemology.* Oxford: Oxford University Press.

Carruthers, P. (2018). Implicit versus explicit attitudes: Differing manifestations of the same representational structures? *Review of Philosophy and Psychology., 9*(1), 52–72.

Forbes, C. E., & Schmader, T. (2010). Retraining attitudes and stereotypes to affect motivation and cognitive capacity under stereotype threat. *Journal of Personality and Social Psychology, 99*(5), 740–754.

Gawronski, B., & Bodenhausen, G. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(2006), 692–731.

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

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

Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70*(3), 491–512.

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

Haslanger, S. (2015). Social structure, narrative, and explanation. *Canadian Journal of Philosophy, 45*(1), 1–15.

Henry, P. J., & Sears, D. O. (2002). The symbolic racism 2000 scale. *Political Psychology, 23*, 2.

Kawakami, K., Dovidio, J. F., & van Kamp, S. (2005). Kicking the habit: Effects of nonstereotypic association training and correction processes on hiring decisions. *Journal of Experimental Social Psychology, 41*(1), 68–75.

Kawakami, K., Dovidio, J. F., & van Kamp, S. (2007). The impact of counter-stereotypic training and related correction processes on the application of stereotypes. *Group Processes and Intergroup Relations, 10*(2), 139–156.
Lebrecht, S., Pierce, L., Tarr, M., & Tanaka, J. (2009). Perceptual other-race training reduces implicit racial bias. *PLoS ONE, 4*(1), 4215.
Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Nous, 49*(4), 800–823.
Levy, N. (2017). Am i a Racist? Implicit bias and the ascription of racism. *The Philosophical Quarterly, 67*, 268.
Machery, E. (2017). Should we throw the IAT on the scrap heap of indirect measures?' Comment on the Brains Blog, January 17. http://philosophyofbrains.com/2017/01/17/how-can-we-measure-implicit-bias-a-brains-blog-roundtable.aspx
Machery, E. (2016). De-freuding implicit attitudes. In J. Saul & M. Brownstein (Eds.), *Implicit bias and philosophy, volume I: Metaphysics and epistemology* (p. 2016). Oxford: Oxford University Press.
Machery, E. (2021). Anomalies in implicit attitudes research. *Wires Cognitive Science. Early View:* https://doi.org/10.1002/wcs.1569
Madva, A. (2017). Biased against debiasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice. *Ergo, 4*(6), 145–179.
Madva, A., & Brownstein, M. (2018). Stereotypies, prejudice, and the taxonomy of the implicit social mind. *Nous, 52*(3), 611–644.
Mandelbaum, E. (2016). Attitude, inference, association: on the propositional structure of implicit bias. *Nous, 50*(3), 629–658.
McConohay, J. (1982). Self-interest versus racial attitudes as correlates of anti-busing attitudes in louisville: Is it the buses or the blacks? *The Journal of Politics, 44*(3), 692–720.
McConohay, J., Hardee, J. B., & Batts, V. (1981). Has racism declined in America? It depends on who is asking and what is asked. *Journal of Conflict Resolution, 25*(4), 563–579.
Nosek, B., & Banaji, M. (2001). The Go/No-Go Association task. *Social Cognition, 19*(6), 625–666.
Oswald, F. L., Mitchell, G., Blanton, H., & Jaccard, J. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Attitudes and Social Cognition, 105*(2), 171–192.
Payne, B. (2009). Attitude misattribution: Implications for attitude measurement and the implicit-explicit relationship. In A. Black, W. Prokasy, R. Petty, R. Fazio, & P. Brinol (Eds.), *Attitudes: Insights from the new wave of implicit measures.* Erlbaum.
Rachlinski, J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2009). Does unconscious racial bias affect trial judges? *Notre Dame Law Review, 84*, 1195.
Sabin, J., Nosek, B., Greenwald, A., & Rivara, F. P. (2009). Physicians' implicit and explicit attitudes about race by MD race, ethnicity and gender. *Journal of Health Care Poor Underserved, 20*(3), 896.
Saul, J. (2012). Scepticism and implicit bias. *Disputatio,* Lecture 2012.
Sie, M., & van Voorst Vader-Bours, N. (2016). "Stereotypes and prejudices: Whose responsibility? Indirect personal responsibility for implicit biases. In M. Brownstein & J. Saul (Eds.), *Philosophy and implicit bias, volume II: Moral responsibility, structural injustice, and ethics.* Oxford University Press.
Singal, J. (2017). Psychology's favorite tool for measuring racism isn't up to the job. *New York Magazine*. https://www.thecut.com/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html
Sriram, N., & Greenwald, A. (2009). Brief implicit association test. *Experimental Psychology, 56*(4), 283–294.
Strawson, P. F. (1962). Freedom and resentment. In *Proceedings of the British Academy* (p. 48).
Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: Old fashioned and modern prejudices. *Journal of Personality and Social Psychology, 68*(2), 199–214.
Washington, N., & Kelly, D. (2016). "Who's responsible for this? Moral responsibility, externalism, and knowledge about implicit bias. In M. Brownstein & J. Saul (Eds.), *Philosophy and implicit bias, volume II: Moral responsibility, structural injustice, and ethics.* Oxford University Press.
Wennekers, A. M., Holland, R. W., Wigboldus, D. H., & van Knippenberg, A. (2012). First see, then nod: The role of temporal contiguity in embodied evaluative conditioning of social attitudes. *Social Psychological and Personality Science, 3*(4), 455–461.
Wittenbreck, 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.
Wolf, S. (2007). Moral psychology and the unity of the virtues. *Ratio, 20*(2), 145–167.

Zheng, R. (2016). Attributability, accountability, and implicit bias. In M. Brownstein & J. Saul (Eds.), *Philosophy and implicit bias, volume II: Moral responsibility, structural injustice, and ethics.* Oxford University Press.