WIREs
Wiley Interdisciplinary Reviews

Article Title: Assessing the Implicit Bias Research Programme: Comments on Brownstein, Gawronski, and Madva vs. Machery

Article Category:

PERSPECTIVE | PRIMER | OVERVIEW
ADVANCED REVIEW | FOCUS ARTICLE | SOFTWARE FOCUS

Authors:

First author full name

Shannon Spaulding, https://orcid.org/0000-0003-4415-479, Oklahoma State University, shannon.spaulding@okstate.edu, no conflict of interest.

Abstract Michael Brownstein, Alex Madva, and Bertram Gawronski (2020, 2022) articulate a careful defense of research on implicit bias. They argue that though there is room for improvement in various areas, when we set the bar appropriately and when we are comparing relevant events, the test-retest stability and predictive ability of implicit bias measures are respectable. Edouard Machery disagrees. He argues that theories of implicit bias have failed to answer four fundamental questions about measures of implicit bias, and this undermines their utility in further scientific research and policy making. In this article, I offer my perspective on this important debate. I argue that there is a theoretical mismatch in debating the merits of a research programme on the terms of a specific theory within the research programme. Nevertheless, the discussion allows us to see which questions are answered from within the

perspective of a particular theory. I argue that the emphasis should be on whether implicit bias theories predict novel facts.

1. INTRODUCTION

The details of particular experiments, meta-analyses, and specific theoretical frameworks are important when mediating debates about measuring implicit bias. Theories of implicit bias make different predictions about when we should see, for example, correlations between different indirect measures of implicit bias, correlations between indirect measures of bias and real-world behavior, and correlations between an individual's performances on indirect measures of bias over time.[i] When we are discussing competing theories of implicit bias that make differing predictions, these details are crucial. However, the debate between Michael Brownstein, Alex Madva, and Bertram Gawronski (2020; 2022) and Edouard Machery (2021, 2022) is not that kind of debate. It is a debate about the legitimacy of the implicit bias research programme itself. In this kind of discussion, the details matter of course, but it is easy to lose sight of the contours of the debate when focusing on the details. Thus, I will keep this short response at a fairly high level of analysis.

[2. RESEARCH PROGRAMMES]

We can think of implicit bias research in terms of Imre Lakatos' (Lakatos, 1978) conception of research programmes. Scientific theories have three components: a *hard core* consisting of central theses of the theory, a *protective belt* of auxiliary assumptions about measuring instruments, environments, etc., and a *positive heuristic* consisting in guidance for how to solve problems using the theory and how to respond to anomalies by revising the auxiliary assumptions. Research programmes consist in competing theories that share the commitments of the hard core but differ in their auxiliary assumptions and positive heuristics. This is the case in the implicit bias research programme. Brownstein, Gawronski, and Madva's theory of implicit bias is just one amongst several competing theories in the implicit bias research programme. Their theory differs from, say, Greenwald, Banaji,

and Nosek, but all the theories share the central commitment that our behavior can reveal subtle but significant biases that are difficult to detect with explicit measures.

Scientific theories all have these three elements, and they all have unsolved problems at any stage of development. Thus, having problems and revising auxiliary assumptions is not unusual or in and of itself problematic for a scientific theory. However, not all research programmes are equally good. *Progressive* research programmes continue to predict novel facts, whereas *degenerating* research programmes do not and fabricate theories only in order to accommodate known facts. On Lakatos' model, theory change occurs as the result of competition between rival research programmes, and scientific revolutions occur when a progressive research programme replaces a degenerating one. Scientists tend to join the progressive research programme, but it is not intellectually dishonest to stick with the degenerating research programme to try to turn it into a progressive one.

The question is how to think of implicit bias research on this model. Are the various theories that constitute the research programme continuing to progress, to predict novel facts? Are theories degenerating, covering up anomalies with post-hoc rationalizations? Is there an alternative research programme out there that solves (or dissolves) some of the fundamental problems that theories of implicit bias have not yet figured out? This is where the real debate between Brownstein, Gawronski, Madva and Machery is.

[3. ASSESSING THE IMPLICIT BIAS RESEARCH PROGRAMME]

Machery objects that 30 years into the research programme, we still do not have decisive answers to four fundamental questions: (1) what indirect tests measure (i.e., whether direct and indirect tests are simply different ways of measuring one thing – attitudes – or whether indirect tests measure *implicit* attitudes), (2) what to make of moderate to low correlations amongst indirect tests, (3) whether results on indirect tests predict real-world behavior, and (4) whether what indirect tests measure is causally efficacious. He takes these four unresolved questions to be *anomalies* in the philosophy of science sense.

One way in which the Lakatosian model helps here is that it makes clear that having unsolved questions is not an indictment of the theory, even long-standing unsolved fundamental questions. For context, evolutionary biologists disagree about whether natural selection operates at the level of genes, individuals, or groups. This open question is at the heart of evolutionary science.[ii] We cannot say for sure what natural selection actually selects, but no one should take this to indicate that evolutionary science is a degenerating or failed research programme. Thus, having unsolved fundamental questions is not a problem in and of itself if, like evolutionary biology, the research programme continues to predict novel facts.

But, what about those unsolved questions? The debate here is between Brownstein, Gawronski, and Madva and Machery. Brownstein, Gawronski, and Madva are not defenders of every theory in the research programme. (Many of these other theories in the research programme have different commitments and predictions – different auxiliary assumptions and positive heuristics in Lakatosian terms – from Brownstein, Gawronski, and Madva's theory.) Thus, it is difficult to give an answer to these questions on behalf of the whole research programme. Nevertheless, Brownstein, Gawronski, and Madva give answers to these questions on behalf of their own theory and argue that these answers stem from longstanding, empirically supported frameworks.

In response to question (1), Brownstein, Gawronski, and Madva argue that indirect and direct tests measure the same thing, namely attitudes. They argue that the dual-attitude or modal interpretation of indirect tests is unwarranted. Questions (2) and (3) requires a nuanced response, and their reply is multi-pronged. First, they emphasize that they interpret responses to indirect tests as task-specific, context-sensitive, person-specific behaviors that operate on momentarily accessible information. My performance on a race IAT differs significantly from another person's performance on a race AMP test, and we should not expect these performances to correlate particularly strongly because of the different tests, different contexts, and different people involved.[iii] The same is true when thinking about behavioral predictions, according to Brownstein, Gawronski, and Madva. Scores on indirect and direct tests will be better at predicting behavior when the examined behavior matches the demands and

context of the experimental test. Thus, when we are searching for meaningful correlations amongst tests and between tests and behavior, we must match apples to apples. (Machery recognizes the need for this kind of precision in matching experimental conditions when discussing Greenwald, Banaji, and Nosek's claim that if police officers' IAT scores were reduced it would result in reduced racial disparities in stops (Machery, 2021, p. p. 7).) Brownstein, Gawronski, and Madva claim that when we do appropriately match tests, the test-retest reliability is at a respectable level and the tests do predict behavior.

Machery worries that the notion of context employed in this last response is so vague that it could license any interpretation of data we like. He asks, "It is also natural to wonder how indirect measure scores could tell us anything of interest about real world behavior since contexts aren't similar when completing an indirect measure of attitudes and in everyday life – and if it happens to be similar to some real-life contexts, it must be different from many others. What then is the relevance of indirect measures for understanding everyday behavior and social ills generally?" (Machery, 2022, p. p. 7). This is an excellent point. It is, to my mind, the most pressing worry about implicit bias research. Brownstein, Gawronski, and Madva do not offer this analogy, but one could look to Whole Trait Theory for a framework for responding to this concern.

According to Whole Trait Theory, character traits are aggregates of a person's behaviors across time and situations (Fleeson & Jayawickreme, 2015). We can model a person's behavior in terms of density distribution of personality states. These states fluctuate from moment to moment and situation to situation, but the aggregate of these states is stable around a central point. For instance, an extrovert may be outgoing in most situations but reserved in particular kinds of situations. The *average* extroversion of their actions remains similar week-to-week. Whole Trait Theory offers a formal model for determining the stability of behavior across time and contexts. Brownstein, Gawronski, and Madva do *not* conceive indirect tests as measuring stable traits. That is why their response to question (4) is that the question does not make sense on their framework. However, they could employ a framework like this to mitigate the worry about context sensitivity undermining predictive utility.

Scaling out a bit, there is a worry about the course of this dialectic. Machery condemns the field for not having answers to these questions; proponents of a theory from the field answer those questions in light of their own theory; Machery points to other theories that have problems or give different answers. The discussants are not quite talking past each other, but it is hard to see how what either side has to say will be satisfying to the other side. It is difficult to debate the merits of a research programme on the terms of a specific theory within the research program. Instead of focusing on specific questions that different theories within a research programme will answer differently, I think a more productive course of inquiry is to ask whether the theories make novel predictions.

Do the theories in the implicit bias research programme make novel predictions? I think the answer is clearly yes. The two articles by Brownstein, Gawronski, and Madva are filled with predictions that flow from their theory that responses to indirect measures are person-specific, situation-specific behaviors that stem from dynamic processes operating on momentarily accessible information. These predictions range from when we should expect results from different kinds of test *not* to correlate well, when the correlations between tests *should* be stronger, how the content of a test will influence its predictive ability, etc. The most pressing concern for Brownstein, Gawronski, and Madva's theory is to figure out how to make specific predictions about behavior in the real world given the complex interaction between person, situation, and context that their theory posits.

Machery is unimpressed by the comment that none of the metaanalyses report nonsignificant correlations close to zero or negative correlations with behavior. And, on the one hand, it is easy to see why one would find this unimpressive when the question is how strong the correlations are. On the other hand, when the question is whether the theories make novel predictions at all, whether these are progressive research programmes, it is a relevant fact. It tells us that the tests are tapping into something that needs explaining (or at least explaining away, for the skeptics).

The final aspect of assessing research programmes concerns whether there is an alternative that fares better, that either solves or dissolves the outstanding questions. For theoretical reasons, this is

a difficult question to answer. Different research programmes will focus on answering different fundamental questions, have different ways of discriminating signal and noise, and make different kinds of testable predictions. We do not need to endorse Kuhnian incommensurability to recognize the difficulty of comparing the merits of different research programmes. These evaluative judgments are easier to make in retrospect than in real-time. Setting aside those theoretical difficulties though, this is an important area for development. In previous work, Machery has articulated a different kind of approach to implicit bias (Machery, 2016). Depending on how finely we discriminate research programmes, Payne et al.'s (2017) work correlating IAT scores and implicit bias at the group-level might count as an alternative research programme. I think it is too early to tell whether these research programmes are better or whether they will, in the end, complement the research programme under discussion here.

[4. SCIENCE COMMUNICATION]

Finally, there is a distinct issue that comes up again and again in Machery's original piece (2021) and his reply (2022). That is the issue of scientific communication of implicit bias research to the public. He regards the oversimplification, overstating what we can know, and overall hype in communicating implicit bias research to the public as extremely problematic. The issue takes up roughly 25% of the original article. Thus, clearly Machery regards this is a relevant and significant critique. Brownstein, Gawronski, and Madva agree that more care needs to be taken in communicating results to the public, but they regard the issue of scientific communication as distinct from whether the implicit bias research programme is progressive.

Both views are right, of course. Interviews and public policy choices have little bearing on whether a research programme predicts novel facts. Fanfare to the public and policy makers does not enhance or detract from the epistemic status of a theory.[iv] However, researchers do have an obligation to not oversell what we can know and do with their research. This obligation is grounded in our social role as experts in a field. It is an obligation that applies to all of us, not just implicit bias researchers. Knowing that scarce time and resources may be invested in implementing our research, experts ought to take

care in communicating their findings. It is clearly true that a great deal of time and money has been spent implementing implicit bias interventions that have little impact on what the interveners want changed, and the researchers who are the public face of this research sometimes oversimplify, oversell, and overpromise when speaking to the public and policy makers. Recent scandals in promoting science (e.g., power posture, Theranos health technology, hydroxychloroquine as a treatment for Covid) highlight the perils of confusing aspirational goals of research with current capabilities. While it is unlikely that implicit bias interventions have harmed anyone (though they may have annoyed or bored some unwilling participants), the time and resources could be put to better use. And, as always, we should be wary of eroding public trust in science by promoting interventions that turn out to be ineffective. This is not a reason to reject the implicit bias research programme, though. It is a reason to do more and better work to figure out how to intervene on biased behavior in the real-world.

References

Brownstein, M., Madva, A., & Gawronski, B. (2020). Understanding implicit bias: Putting the criticism into perspective. *Pacific Philosophical Quarterly*.

Del Pinal, G., & Spaulding, S. (2018). Conceptual centrality and implicit bias. *Mind & Language, 33*(1), 95-111.

Fleeson, W., & Jayawickreme, E. (2015). Whole Trait Theory. *Journal of Research in Personality, 56*, 82-92. doi:https://doi.org/10.1016/j.jrp.2014.10.009

Gawronski, B., Brownstein, M., & Madva, A. (2022). How Should We Think About Implicit Measures and Their Empirical "Anomalies"? *Wiley Interdisciplinary Reviews: Cognitive Science*.

Lakatos, I. (1978). Science and pseudoscience. *Philosophical papers, 1*, 1-7.

Machery, E. (2016). De-Freuding implicit attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit Bias & Philosophy* (Vol. 1). Oxford: Oxford University Press.

Machery, E. (2021). Anomalies in implicit attitudes research. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1569.

Machery, E. (2022). Anomalies in Implicit Attitude Research: Not So Easily Dismissed. *Wiley Interdisciplinary Reviews: Cognitive Science*.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*(4), 233-248.

[i] I will follow Machery in distinguishing indirect/direct *tests* from implicit/explicit *measurands*.

[ii] Of course, some theories are committed to one unit of analysis or another, and some theorists would not regard this as an open question anymore. This highlights a thorny issue of when to characterize a question as unresolved. The mere fact of disagreement (even disagreement amongst qualified experts) is too low of a bar. The standard cannot be unanimous agreement amongst experts because that almost never occurs. But then what is the standard for characterizing a question as unresolved? I do not know how to set a principled threshold for counting some level of disagreement as indicative of an unresolved question. This matters because Brownstein, Gawronski, and Madva could reasonably say that they do not regard some of these questions as unresolved. They can, and do, maintain that there are theoretically motivated and empirically

supported answers to some of the questions that Machery characterizes as unresolved. I do not want to rest too much on this point, but it is worth remarking that it can be a bit of a shaky foundation for critique.

[iii] The debate here is between Machery and Brownstein, Gawronski, and Madva, but for interested readers I will point out that in previous work Guillermo Del Pinal and I have given independent reasons to expect such correlations across different kinds of tests to be weak. See Del Pinal and Spaulding (2018).

[iv] Though, perhaps fanfare generates more research funding, which generates more studies that can be used to test and develop the theory.