Scientism in the Philosophy of Implicit Bias Research

Kamili Posey, Kingsborough College


Short url: https://wp.me/p1Bfg0-41m (provided by WordPress)
If you consider the recent philosophical literature on implicit bias research, then you would be forgiven for thinking that the problem of successful interventions into implicit bias fall into the category of things that are resolved. If you consider the recent social psychological literature on interventions into implicit bias, then you would come away with a similar impression. The claim is that implicit bias is epistemically harmful because we profess to believing one thing while our implicit attitudes tell a different story.

**Strategy Models and Discrepancy Models**

Implicit bias is socially harmful because it maps onto our real-world discriminatory practices, e.g., workplace discrimination, health disparities, racist police shootings, and identity-prejudicial public policies. Consider the results of Greenwald et al.’s (1998) Implicit Association Test. Consider also the results of Correll et. al’s (2002) “Shooter Bias.” If cognitive interventions are possible, and specifically implicit cognitive interventions, then they can help knowers implicitly manage automatic stereotype activation. Do these interventions lead to real-world reductions of bias?

Linda Alcoff (2010) notes that it is difficult to see how implicit, nonvolitional biases (e.g., those at the root of social and epistemic ills like race-based police shootings) can be remedied by explicit epistemic practices. I would follow this by noting that it is equally difficult to see how nonvolitional biases can be remedied by implicit epistemic practices as well.

Jennifer Saul (2017) responds to Alcoff’s (2010) query by pointing to social psychological experiments conducted by Margo Monteith (1993), Jack Glaser and Eric D. Knowles (2007), Gordon B. Moskowitz and Peizhong Li (2011), Saaid A. Mendoza et al. (2010), Irene V. Blair et al. (2001), and Kerry Kawakami et al. (2005). These studies suggest that implicit self-regulation of implicit bias is possible. Saul notes that philosophers with objections like Alcoff’s, and presumably like mine, should “not just to reflect upon the problem from the armchair - at the very least, one should use one’s laptop to explore the internet for effective interventions.”

But I think this recrimination rings rather hollow. How entitled are we to extrapolate from social psychological studies in the manner that Saul advocates? How entitled are we to assumes the epistemic superiority of scientific research on racism, sexism, etc. over the phenomenological reporting of marginalized knowers? Lastly, how entitled are we to claims about the real-world applicability of these study results? My guess is that the devil is in the

---

details. My guess is also that social psychologists have not found the silver bullet for remedying implicit bias. But let’s follow Saul’s suggestion and not just reflect from the armchair.

A caveat: the following analysis is not intended to be an exhaustive or thorough refutation of what is ultimately a large body social psychological literature. Instead, it is intended to cast a bit of doubt on how these models are used by philosophers as successful remedies for implicit bias. It is intended to cast doubt a bit of doubt on the idea that remedies for racist, sexist, homophobic, and transphobic discrimination are merely a training session or reflective exercise away.

This type of thinking devalues the very real experiences of those who live through racism, sexism, homophobia, and transphobia. It devalues how pervasive these experiences are in American society and the myriad ways in which the effects of discrimination seep into marrow of marginalized bodies and marginalized communities. Worse still, it implies that marginalized knowers who claim, “You don’t understand my experiences!” are compelled to contend with the hegemonic role of “Science” that continues to speak over their own voices and about their own lives. But again, back to the studies.

**Four Methods of Remedy**

I break up the above studies into four intuitive model types: (1) strategy models, (2) discrepancy models, (3) (IAT) models, and (4) egalitarian goal models. (I am not a social scientist, so the operative word here is “intuitive.”) Let’s first consider Kawakami et al. (2005) and Mendoza et al. (2010) as examples of strategy models. Kawakami et al. used Devine and Monteith’s (1993) notion of a negative stereotype as a “bad habit” that a knower needs to “kick” to model strategies that aid in the inhibition of automatic stereotype activation, or the inhibition of “increased cognitive accessibility of characteristics associated with a particular group.”

---

5 I owe this critical point in its entirety to the work of Lacey Davidson and her presentation, “When Testimony Isn’t Enough: Implicit Bias Research as Epistemic Injustice” at the Feminist Epistemologies, Methodologies, Metaphysics, and Science Studies (FEMMSS) conference in Corvallis, Oregon in 2018. Davidson notes that the work of philosophers of race and critical race theorists often takes a backseat to the projects of philosophers of social science who engage with the science of racialized attitudes as opposed to the narratives and/or testimonies of those with lived experiences of racism. Davidson describes this as a type of epistemic injustice against philosophers of race and critical race theorists. She also notes that philosophers of race and critical race theorists are often people of color while the philosophers of social science are often white. This dimension of analysis is important but unexplored. Davidson’s work highlights how epistemic injustice operates within the academy to perpetuate systems of racism and oppression under the guise of “good science.” Her arguments was inspired by the work of Jeanine Weckes Schroer on the problematic nature of current research on stereotype threat and implicit bias in “Giving Them Something They Can Feel: On the Strategy of Scientizing the Phenomenology of Race and Racism,” Knowledge Cultures 3(1), 2015.

In a previous study, Kawakami et al. (2000) asked research participants presented with photographs of black individuals and white individuals with stereotypical traits and non-stereotypical traits listed under each photograph to respond “No” to stereotypical traits and “Yes” to non-stereotypical traits. The study found that “participants who were extensively trained to negate racial stereotypes initially also demonstrated stereotype activation, this effect was eliminated by the extensive training.

Furthermore, Kawakami et al. found that practice effects of this type lasted up to 24 h following the training.” Kawakami et al. (2005) used this training model to ground an experiment aimed at strategies for reducing stereotype activation in the preference of men over women for leadership roles in managerial positions. Despite the training, they found that there was “no difference between Nonstereotypic Association Training and No Training conditions...participants were indeed attempting to choose the best candidate overall, in these conditions there was an overall pattern of discrimination against women relative to men in recommended hiring for a managerial position (Glick, 1991; Rudman & Glick, 1999)” [emphasis mine].

Substantive conclusions are difficult to make by a single study but one critical point is how learning occurred in the training but improved stereotype inhibition did not occur. What, exactly, are we to make of this result? Kawakami et al. (2005) claimed that “similar levels of bias in both the Training and No Training conditions implicates the influence of correction processes that limit the effectiveness of training.” That is, they attributed the lack of influence of corrective processes on a variety of contributing factors that limited the effectiveness of the strategy itself.

Notice, however, that this does not implicate the strategy as a failed one. Most notably Kawakami et al. found that “when people have the time and opportunity to control their responses [they] may be strongly shaped by personal values and temporary motivations, strategies aimed at changing the automatic activation of stereotypes will not [necessarily] result in reduced discrimination.”

This suggests that although the strategies failed to reduce stereotype activation they may still be helpful in limited circumstances “when impressions are more deliberative.” One wonders under what conditions such impressions can be more deliberative? More than that, how useful are such limited-condition strategies for dealing with everyday life and every day automatic stereotype activation?

---

8 Kawakami et al. (2005), p. 69.
9 Kawakami et al. (2005), p. 73.
10 Kawakami et al. (2005), p. 73.
11 Kawakami et al. (2005), p. 74.
12 Kawakami et al. (2005), p. 74.
Mendoza et al. (2010) tested the effectiveness of “implementation intentions” as a strategy to reduce the activation or expression of implicit stereotypes using the Shooter Task. They tested both “distraction-inhibiting” implementation intentions and “response-facilitating” implementation intentions. Distraction-inhibiting intentions are strategies “designed to engage inhibitory control,” such as inhibiting the perception of distracting or biasing information, while “response-facilitating” intentions are strategies designed to enhance goal attainment by focusing on specific goal-directed actions.

In the first study, Mendoza et al. asked participants to repeat the on-screen phrase, “If I see a person, then I will ignore his race!” in their heads and then type the phrase into the computer. This resulted in study participants having a reduced number of errors in the Shooter Task. But let’s come back to if and how we might be able to extrapolate from these results. The second study compared a simple-goal strategy with an implementation intention strategy.

Study participants in the simple-goal strategy group were asked to follow the strategy, “I will always shoot a person I see with a gun!” and “I will never shoot a person I see with an object!” Study participants in the implementation intention strategy group were asked to use a conditional, if-then, strategy instead: “If I see a person with an object, then I will not shoot!” Mendoza et al. found that a response-facilitating implementation intention “enhanced controlled processing but did not affect automatic stereotyping processing,” while a distraction-inhibiting implementation intention “was associated with an increase in controlled processing and a decrease in automatic stereotyping processes.”

How to Change Both Action and Thought

Notice that if the goal is to reduce automatic stereotype activation through reflexive control that only a distraction-inhibiting strategy achieved the desired effect. Notice also how the successful use of a distraction-inhibiting strategy may require a type of “non-messy” social environment unachievable outside of a laboratory experiment. Or, as Mendoza et al. (2010)
rightly note: “The current findings suggest that the quick interventions typically used in psychological experiments may be more effective in modulating behavioral responses or the temporary accessibility of stereotypes than in undoing highly edified knowledge structures.”

The hope, of course, is that distraction-inhibiting strategies can help dominant knowers reduce automatic stereotype activation and response-facilitated strategies can help dominant knowers internalize controlled processing such that negative bias and stereotyping can be (one day) reflexively controlled as well. But these are only hopes. The only thing that we can rightly conclude from these results is that if we ask a dominant knower to focus on an internal command, they will do so. The result is that the activation of negative bias fails to occur.

This does not mean that the knower has reduced their internalized negative biases and prejudices or that they can continue to act on the internal commands in the future (in fact, subsequent studies reveal the effects are short-lived). As Mendoza et al. also note: “In psychometric terms, these strategies are designed to enhance accuracy without necessarily affecting bias. That is, a person may still have a tendency to associate Black people with violence and thus be more likely to shoot unarmed Blacks than to shoot unarmed Whites.”

Despite hope for these strategies, there is very little to support their real-world applicability.

Hunting for Intuitive Hypocrisies

I would extend a similar critique to Margot Monteith’s (1993) discrepancy model. Monteith’s (1993) often cited study uses two experiments to investigate prejudice related discrepancies in the behaviors of low-prejudice (LP) and high-prejudice (HP) individuals and the ability to engage in self-regulated prejudice reduction. In the first experiment, (LP) and (HP) heterosexual study participants were asked to evaluate two law school applications, one for an implied gay applicant and one for an implied heterosexual applicant. Study participants “were led to believe that they had evaluated a gay law school applicant negatively because of his sexual orientation;” they were tricked into a “discrepancy-activated condition” or a condition that was at odds with their believed prejudicial state.

All of the study participants were then told that the applications were identical and that those who had rejected the gay applicant had done so because of the applicant’s sexual orientation.

are predicated on the noninterference of other destabilizing factors, be meaningfully applied to everyday life?

There is a tendency in the philosophical literature on implicit bias and stereotype threat to outright ignore the limited applicability of much of this research in order to make critical claims about interventions into racist, sexist, homophobic, and transphobic behaviors. Philosophers would do well to recognize the complexity of these issues and to be more cautious about the enthusiastic endorsement of experimental results.

17 Mendoza, Saaid, Gollwitzer, Peter, and Amodio, David (2010), p. 520.
19 Mendoza, Saaid, Gollwitzer, Peter, and Amodio, David (2010), p. 520.
It is important to note that the applicants qualifications were not, in fact, identical. The gay applicant’s application materials were made to look worse than the heterosexual applicant’s materials. This was done to compel the rejection of the applicant.

Study participants were then provided a follow-up questionnaire and essay allegedly written by a professor who wanted to know (a) “why people often have difficulty avoiding negative responses toward gay men,” and (b) “how people can eliminate their negative responses toward gay men.” Researchers asked study participants to record their reactions to the faculty essay and write down as much they could remember about what they read. They were then told about the deception in the experiment and told why such deception was incorporated into the study.

Monteith (1993) found that “low and high prejudiced subjects alike experienced discomfort after violating their personal standards for responding to a gay man, but only low prejudiced subjects experienced negative self-directed affect.” Low prejudiced, (LP), “discrepancy-activated subjects,” also spent more time reading the faculty essay and “showed superior recall for the portion of the essay concerning why prejudice-related discrepancies arise.”

The “discrepancy experience” generated negative self-directed affect, or guilt, for (LP) study participants with the hope that the guilt would (a) “motivate discrepancy reduction (e.g., Rokeach, 1973)” and (b) “serve to establish strong cues for punishment (cf. Gray, 1982).” The idea here is that the experiment results point to the existence of a self-regulatory mechanism that can replace automatic stereotype activation with “belief-based responses;” however, “it is important to note that the initiation of self-regulatory mechanisms is dependent on recognizing and interpreting one’s responses as discrepant from one’s personal beliefs.”

The discrepancy between what one is shown to believe and what one professes to believe (whether real or manufactured, as in the experiment) is aimed at getting knowers to engage in heightened self-focus due to negative self-directed affect. The goal of Monteith’s (1993) study is that self-directed affect would lead to a kind of corrective belief-making process that is both less prejudicial and future-directed.

But if it’s guilt that’s doing the psychological work in these cases, then it’s not clear that knowers wouldn’t find other means of assuaging such feelings. Why wouldn’t it be the case that generating negative self-directed affect would point a knower toward anything they deem necessary to restore a more positive sense of self? To this, Monteith made the following concession:

---

Steele (1988; Steele & Liu, 1983) contended that restoration of one’s self-image after a discrepancy experience may not entail discrepancy reduction if other opportunities for self-affirmation are available. For example, Steele (1988) suggested that a smoker who wants to quit might spend more time with his or her children to resolve the threat to the self-concept engendered by the psychological inconsistency created by smoking. Similarly, Tesser and Cornell (1991) found that different behaviors appeared to feed into a general “self-evaluation reservoir.” It follows that prejudice-related discrepancy experiences may not facilitate the self-regulation of prejudiced responses if other means to restoring one’s self-regard are available [emphasis mine].

Additionally, she noted that even if individuals are committed to the reducing or “unlearning” automatic stereotyping, they “may become frustrated and disengage from the self-regulatory cycle, abandoning their goal to eliminate prejudice-like responses.” Cognitive exhaustion, or cognitive depletion, can occur after intergroup exchanges as well. This may make it even less likely that a knower will continue to feel guilty, and to use that guilt to inhibit the activation of negative stereotypes when they find themselves struggling cognitively. Conversely, there is also the issue of a kind of lab-based, or experiment-based, cognitive priming. I pick up with this idea along with the final two models of implicit interventions in the next part.

(IAT) Models and Egalitarian Goal Models

In this half, I go through the last two models, Glaser and Knowles’ (2007) and Blair et al.’s (2001) (IAT) models and Moskowitz and Li’s (2011) egalitarian goal model. I reiterate that this is not an exhaustive analysis of such models nor is it intended as a criticism of experiments pertaining to implicit bias. Mostly, I am concerned that the science is interesting but that the scientism - the application of tentative results to philosophical projects - is less so. It is from this point that I proceed.

Like Mendoza et al.’s (2010) implementation intentions, Glaser and Knowles’ (2007) (IMCP) aims to capture implicit motivations that are capable of inhibiting automatic stereotype activation. Glaser and Knowles measure (IMCP) in terms of an implicit negative attitude toward prejudice, or (NAP), and an implicit belief that oneself is prejudiced, or (BOP). This is done by retooling the (IAT) to fit both (NAP) and (BOP): “To measure NAP we constructed an IAT that pairs the categories ‘prejudice’ and ‘tolerance’ with the categories ‘bad’ and ‘good.’ BOP was assessed with an IAT pairing ‘prejudiced’ and ‘tolerant’ with ‘me’ and ‘not me.’”

Study participants were then administered the Shooter Task, the (IMCP) measures, and the Race Prejudice (IAT) and Race-Weapons Stereotype (RWS) tests in a fixed order. They predicted that (IMCP) as an implicit goal for those high in (IMCP) “should be able to short-

circuit the effect of implicit anti-Black stereotypes on automatic anti-Black behavior."\(^{29}\) The results seemed to suggest that this was the case. Glaser and Knowles found that study participants who viewed prejudice as particularly bad "[showed] no relationship between implicit stereotypes and spontaneous behavior."\(^{30}\)

There are a few considerations missing from the evaluation of the study results. First, with regard to the Shooter Task, Glaser and Knowles (2007) found that "the interaction of target race by object type, reflecting the Shooter Bias, was not statistically significant."\(^{31}\) That is, the strength of the relationship that Correll et al. (2002) found between study participants and the (high) likelihood that they would "shoot" at black targets was not found in the present study. Additionally, they note that they "eliminated time pressure" from the task itself. Although it was not suggested that this impacted the usefulness of the measure of Shooter Bias, it is difficult to imagine that it did not do so. To this, they footnote the following caveat:

Variance in the degree and direction of the stereotype endorsement points to one reason for our failure to replicate Correll et. al’s (2002) typically robust Shooter Bias effect. That is, our sample appears to have held stereotypes linking Blacks and weapons/aggression/danger to a lesser extent than did Correll and colleagues’ participants. In Correll et al. (2002, 2003), participants one \(SD\) below the mean on the stereotype measure reported an anti-Black stereotype, whereas similarly low scorers on our RWS IAT evidenced a stronger association between \(Whites\) and weapons. Further, the adaptation of the Shooter Task reported here may have been less sensitive than the procedure developed by Correll and colleagues. In the service of shortening and simplifying the task, we used fewer trials, eliminated time pressure and rewards for speed and accuracy, and presented only one background per trial.\(^{32}\)

Glaser and Knowles claimed that the interaction of the (RWS) with the Shooter Task results proved "significant," however, if the Shooter Bias failed to materialize (in the standard Correll et al. way) with study participants, it is difficult to see how the (RWS) was measuring anything except itself, generally speaking. This is further complicated by the fact that the interaction between the Shooter Bias and the (RWS) revealed "a mild reverse stereotype associating Whites with weapons \((d = -0.15)\) and a strong stereotype associating Blacks with weapons \((d = 0.83)\), respectively."\(^{33}\)

Recall that Glaser and Knowles (2007) aimed to show that participants high in (IMCP) would be able to inhibit implicit anti-black stereotypes and thus inhibit automatic anti-black behaviors. Using (NAP) and (BOP) as proxies for implicit control, participants high in (NAP) and moderate in (BOP) - as those with moderate (BOP) will be motivated to avoid bias - should show the weakest association between (RWS) and Shooter Bias. Instead, the

lowest levels of Shooter Bias were seen in “low NAP, high BOP, and low RWS” study participants, or those who do not disapprove of prejudice, would describe themselves as prejudiced, and also showed lowest levels of (RWS).³⁴

They noted that neither “NAP nor BOP alone was significantly related to the Shooter Bias,” but “the influence of RWS on Shooter Bias remained significant.”³⁵ In fact, greater bias was actually found with higher (NAP) and (BOP) levels.³⁶ This bias seemed to map on to the initial results of the Shooter Task results. It is most likely that (RWS) was the most important measure in this study for assessing implicit bias, not, as the study claimed, for assessing implicit motivation to control prejudice.

What Kind of Bias?

It is also not clear that the (RWS) was not capturing explicit bias instead of implicit bias in this study. At the point at which study participants were tasked with the (RWS), automatic stereotype activation may have been inhibited just in virtue of study participants involvement in the Shooter Task and (IAT) assessments regarding race-related prejudice. That is, race-sensitivity was brought to consciousness in the sequencing of the test process.

Although we cannot get into the heads of the study participants, this counter explanation seems a compelling possibility. That is, that the sequential tasks involved in the study captured study participants’ ability to increase focus and increase conscious attention to the race-related (IAT) test. Additionally, it is possible that some study participants could both cue and follow their own conscious internal commands, “If I see a black face, I won’t judge!” Consider that this is exactly how implementation intentions work.

Consider that this is also how Armageddon chess and other speed strategy games work. In Park et al.’s (2008) follow-up study on (IMCP) and cognitive depletion, they retreat somewhat from their initial claims about the implicit nature of (IMCP):

> We cannot state for certain that our measure of IMCP reflects a purely nonconscious construct, nor that differential speed to “shoot” Black armed men vs. White armed men in a computer simulation reflects purely automatic processes. Most likely, the underlying stereotypes, goals, and behavioral responses represent a blend of conscious and nonconscious influences...Based on the results of the present study and those of Glaser and Knowles (2008), it would be premature to conclude that IMCP is a purely and wholly automatic construct, meeting the “four horsermen” criteria (Bargh, 1990). Specifically, it is not yet clear whether high IMCP participants initiate control of prejudice without intention; whether implicit control of prejudice

³⁴ Glaser, Jack and Knowles, Eric D. (2007), p. 169. Of this “rogue” group, Glaser and Knowles note: “This group had, on average, a negative RWS (i.e., rather than just a low bias toward Blacks, they tended to associate Whites more than Blacks with weapons; see footnote 4). If these reversed stereotypes are also uninhibited, they should yield reversed Shooter Bias, as observed here” (169).
can itself be inhibited, if for some reason someone wanted to; nor whether IMCP-instigated control of spontaneous bias occurs without awareness.\textsuperscript{37}

If the (IMCP) potentially measures low-level conscious attention, this makes the question of what implicit measurements \textit{actually} measure in the context of sequential tasks all the more important. In the two final examples, Blair et al.’s (2001) study on the use of counterstereotype imagery and Moskowitz and Li’s (2011) study on the use of counterstereotype egalitarian goals, we are again confronted with the issue of sequencing. In the study by Moskowitz and Li, study participants were asked to write down an example of a time when “they failed to live up to the ideal specified by an egalitarian goal, and to do so by relaying an event relating to African American men.”\textsuperscript{38}

They were then given a series of computerized LDTs (lexicon decision tasks) and primes involving photographs of black and white faces and stereotypical and non-stereotypical attributes of black people (crime, lazy, stupid, nervous, indifferent, nosy). Over a series of four experiments, Moskowitz and Li found that when egalitarian goals were “accessible,” study participants were able to successfully generate stereotype inhibition. Blair et al. asked study participants to use counterstereotypical (CS) gender imagery over a series of five experiments, e.g., “Think of a strong, capable woman,” and then administered a series of implicit measures, including the (IAT).

Similar to Moskowitz and Li (2011), Blair et al. (2001) found that (CS) gender imagery was successful in reducing implicit gender stereotypes leaving “little doubt that the CS mental imagery per se was responsible for diminishing implicit stereotypes.”\textsuperscript{39} In both cases, the study participants were explicitly called upon to focus their attention on experiences and imagery pertaining to negative stereotypes before the implicit measures, i.e., tasks, were administered. Again it is not clear that the implicit measures measured the supposed target.

In the case of Moskowitz and Li’s (2011) experiment, the study participants began by relating moments in their lives where they failed to live up to their goals. However, those goals can only be understood within a particular social and political framework where holding negatively prejudicial beliefs about African-American men is often \textit{explicitly} judged harshly, even if not implicitly so. Given this, we might assume that the study participants were compelled into a negative affective state. But does this matter? As suggested by the study by Monteith (1993), and later study by Amodio et. al (2007), guilt can be a powerful tool.\textsuperscript{40}


Questions of Guilt

If guilt was produced during the early stages of the experiment, it may have also participated in the inhibition of stereotype activation. Moskowitz and Li (2011) noted that “during targeted questioning in the debriefing, no participants expressed any conscious intent to inhibit stereotypes on the task, nor saw any of the tasks performed during the computerized portion of the experiment as related to the egalitarian goals they had undermined earlier in the session.”

But guilt does not have to be conscious for it to produce effects. The guilt produced by recalling a moment of negative bias could be part and parcel of a larger feeling of moral failure. Moskowitz and Li needed to adequately disambiguate competing implicit motivations for stereotype inhibition before arriving at a definitive conclusion. This, I think, is a limitation of the study.

However, the same case could be made for (CS) imagery. Blair et al. (2001) noted that it is, in fact, possible that they too have missed competing motivations and competing explanations for stereotype inhibition. Particularly, they suggested that by emphasizing counterstereotyping the researchers “may have communicated the importance of avoiding stereotypes and increased their motivation to do so.” Still, the researchers dismissed that this would lead to better (faster, more accurate) performance of the (IAT), but that is merely asserting that the (IAT) must measure exactly what the (IAT) claims that it does. Fast, accurate, and conscious measures are excluded from that claim. Complicated internal motivations are excluded from that claim.

But on what grounds? Consider Fielder et al.’s (2006) argument that the (IAT) is susceptible to faking and strategic processing, or Brendl et al.’s (2001) argument that it is not possible to infer a single cause from (IAT) results, or Fazio and Olson’s (2003) claim “the IAT has little to do with what is automatically activated in response to a given stimulus.”

These studies call into question the claim that implicit measures like the (IAT) can measure implicit bias in the clear, problem-free manner that is often suggested in the literature. Implicit interventions into implicit bias that utilize the (IAT) are difficult to support for this reason. Implicit interventions that utilize sequential (IAT) tasks are also difficult to support for this reason. Of course, this is also live debate and the problems I have discussed here are far from the only ones that plague this type of research.

44 There is significant debate over the issue of whether the implicit bias that (IAT) tests measure translate into real-world discriminatory behavior. This is a complex and compelling issue. It is also an issue that could render
That said, when it comes to this research we are too often left wondering if the measure itself is measuring the right thing. Are we capturing implicit bias or some other socially generated phenomenon? Are the measured changes we see in study results reflecting the validity of the instrument or the cognitive maneuverings of study participants? These are all critical questions that need sussing out. The temporary result is that the target conclusion that implicit interventions will lead to reductions in real-world discrimination will move further away.\textsuperscript{15} We find evidence of this conclusion in Forscher et al.’s (2018) meta-analysis of 492 implicit interventions:

We found little evidence that changes in implicit measures translated into changes in explicit measures and behavior, and we observed limitations in the evidence base for implicit malleability and change. These results produce a challenge for practitioners who seek to address problems that are presumed to be caused by automatically retrieved associations, as there was little evidence showing that change in implicit measures will result in changes for explicit measures or behavior...Our results suggest that current interventions that attempt to change implicit measures will not consistently change behavior in these domains. These results also produce a challenge for researchers who seek to understand the nature of human cognition because they raise new questions about the causal role of automatically retrieved associations...To better understand what the results mean, future research should

\begin{flushleft}
\begin{minipage}{\textwidth}
moot the (IAT) as an implicit measure of anything full stop. Anthony G. Greenwald, Mahzarin R. Banaji, and Brian A. Nosek (2015) write: “IAT measures have two properties that render them problematic to use to classify persons as likely to engage in discrimination. Those two properties are modest test–retest reliability (for the IAT, typically between $r = .5$ and $r = .6$; cf., Nosek et al., 2007) and small to moderate predictive validity effect sizes. Therefore, attempts to diagnostically use such measures for individuals risk undesirably high rates of erroneous classifications. These problems of limited test-retest reliability and small effect sizes are maximal when the sample consists of a single person (i.e., for individual diagnostic use), but they diminish substantially as sample size increases. Therefore, limited reliability and small to moderate effect sizes are not problematic in diagnosing system-level discrimination, for which analyses often involve large samples” (557). However, Oswald et al. (2013) argue that “IAT scores correlated strongly with measures of brain activity but relatively weakly with all other criterion measures in the race domain and weakly with all criterion measures in the ethnicity domain. IATs, whether they were designed to tap into implicit prejudice or implicit stereotypes, were typically poor predictors of the types of behavior, judgments, or decisions that have been studied as instances of discrimination, regardless of how subtle, spontaneous, controlled, or deliberate they were. Explicit measures of bias were also, on average, weak predictors of criteria in the studies covered by this meta-analysis, but explicit measures performed no worse than, and sometimes better than, the IATs for predictions of policy preferences, interpersonal behavior, person perceptions, reaction times, and microbehavior. Only for brain activity were correlations higher for IATs than for explicit measures...but few studies examined prediction of brain activity using explicit measures. Any distinction between the IATs and explicit measures is a distinction that makes little difference, because both of these means of measuring attitudes resulted in poor prediction of racial and ethnic discrimination” (182-183). For further details about this debate, see: Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P. E. (2013). “Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies,” in Journal of Personality and Social Psychology, Vol. 105, pp. 171-192 and Greenwald, Anthony G., Banaji, Mahzarin R., and Nosek, Brian A. (2015). “Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects,” in Journal of Personality and Social Psychology, Vol. 108, No. 4, pp. 553-561.

\end{minipage}
\end{flushleft}
innovate with more reliable and valid implicit, explicit, and behavioral tasks, intensive manipulations, longitudinal measurement of outcomes, heterogeneous samples, and diverse topics of study.46

Finally, what I take to be behind Alcoff’s (2010) critical question at the beginning of this piece is a kind of skepticism about how individuals can successfully tackle implicit bias through either explicit or implicit practices without the support of the social spaces, communities, and institutions that give shape to our social lives. Implicit bias is related to the culture one is in and the stereotypes it produces. So instead of insisting on changing people to reduce stereotyping, what if we insisted on changing the culture?

As Alcoff notes: “We must be willing to explore more mechanisms for redress, such as extensive educational reform, more serious projects of affirmative action, and curricular mandates that would help to correct the identity prejudices built up out of faulty narratives of history.”47 This is an important point. It is a point that philosophers who work on implicit bias would do well to take seriously.

Science may not give us the way out of racism, sexism, and gender discrimination. At the moment, it may only give us tools for seeing ourselves a bit more clearly. Further claims about implicit interventions appear as willful scientism. They reinforce the belief that science can cure all of our social and political ills. But this is magical thinking.

Contact details: Kamili.Posey@kbcc.cuny.edu

References


