PSYCHOLOGICAL INQUIRY
2022, VOL. 33, NO. 3, 162–166
https://doi.org/10.1080/1047840X.2022.2106753

Routledge
Taylor & Francis Group

COMMENTARIES

So close, Yet So Far: Stopping Short of Killing Implicit Bias

Joseph Cesario

Michigan State University, East Lansing, Michigan

The authors of the target article (Gawronski, Ledgerwood, & Eastwick, this issue) are to be commended for their important and insightful analysis on the state of implicit bias research. They introduce and discuss the critical distinction between bias on implicit measures and implicit bias itself. However, the authors want to have their cake and eat it too, and this causes them to stop short in fully applying their analysis. In this commentary, I take the authors seriously and draw out their analysis to its logical conclusion. In doing so, three points are raised:

1. The central distinction made by the authors, if correct, undermines the claimed importance of implicit bias that researchers have been selling to the public for nearly three decades.
2. The authors sidestep crucial questions about the real-world application of implicit bias that must be answered, given their stated definitions.
3. Researchers risk repeating the failed research trajectory of implicit bias with their newest grand idea, "systemic bias."

After addressing these three main points in detail, I end with a few thoughts on the authors' definition of implicit bias.

What Is the Evidence for Lack of Awareness?

For the sake of argument, let us agree with the authors that the best definition of implicit bias is that of categorical effects without awareness. The important question that follows is, What is the evidence for such effects? Indeed, this is a fundamental question that researchers have now had decades to address, as I described in a recent critique (Cesario, 2022):

> A first-order, foundational question then is whether people are aware of their biases, aware of what is being assessed during the measurement of these biases, or aware of the effects of their biases. After all, if one defines implicit bias as discrimination based on "unconscious" processes and argues that implicit bias is so important as to have implications for legal doctrine in the United States (Greenwald & Krieger, 2006; Kang & Banaji, 2006), then certainly the basic question of awareness must have been thoroughly settled by now … It is striking that the concept of implicit bias has been pushed into federal policy at the highest levels of the U.S. government without any convincing evidence concerning even basic questions about the measurement or the effects of implicit bias.

Clearly my assessment is that implicit bias researchers have not produced the necessary evidence required for their claims. But one does not have to take my word for it. Here is the first author of the target article himself (Gawronski, 2019):

> there is currently no evidence that people are unaware of the mental contents underlying their responses on implicit measures … the preliminary evidence that implicit, but not explicit, biases influence judgment outside of awareness is rather weak and prone to alternative interpretations.

And again even within in the target article itself:

> studies on the relation between BIM and behavior rarely include appropriate awareness checks to confirm the unconscious nature of the effects of social category cues on the focal behavior … (Gawronski et al., this issue, p. 145)

Rather than concluding that the evidence for implicit bias is shockingly weak and that the concept should never have been exported to the public, the authors of the target article instead double-down on the concept and go so far as to claim that "it seems likely that unconscious effects of social category cues contribute to disparities at the social level" (Gawronski et al., this issue, p. 149).

On what possible basis could the authors make such a claim? The answer is that this is pure speculation. Attributing group disparities to implicit bias is wishful thinking.

The authors argue (without evidence) that experimental demonstrations of implicit bias can inform societal-level disparities because "If consequential decisions … are influenced by social category cues … at a sufficiently high rate, they will surely lead to systematic disparities at the societal level" (Gawronski et al., this issue, p. 149). This reasoning reflects several crucial errors commonly made by researchers when connecting experimental demonstrations of bias to real-world decision scenarios, errors which I outlined in a previous paper (Cesario, 2022) familiar to at least the second author (Ledgerwood et al., 2022). In fact, the critical assumptions necessary to connect the two were detailed over twenty years ago within the economics literature on "audit studies" (Heckman, 1998), in which job applications with different demographic information are sent to prospective employers as a means of detecting bias. Without firmly establishing the soundness of these underlying assumptions, which to my knowledge has never done by social psychologists, *experimental demonstrations of bias simply cannot be used to explain real-world disparities.*

Again I ask: What exactly is the causal evidence that group disparities are the result of implicit biases? In

CONTACT Joseph Cesario cesario@msu.edu Michigan State University, East Lansing, Michigan, USA.

response to an article that raised this exact question (Cesario, 2022), a number of implicit bias scholars wrote commentaries defending the claim that implicit bias explains real-world disparities, citing many studies supposedly linking the two. In a reply, I thoroughly evaluated every study and showed that *not a single one* provided any conclusive (or even reasonably strong) evidence on the specific claim that people's unknown categorical bias leads to group disparities. This is simple to see even in the experimental context as almost *no* researchers adequately test for awareness, a fact acknowledged in the target article (see also Gawronski, 2019). *Hence there cannot be strong evidence for implicit bias to explain societal disparities* if there is not even convincing experimental evidence for the foundational claims.

Research often cited as supportive of implicit bias is *in principle* unlikely to fill this role. In the same way that participants taking the IAT become aware of how the categories are influencing their responses, the same is true for behavioral tasks in social cognition that rely on within-subjects manipulations. The First-Person Shooter Task, for example, is designed to study the effects of race on the decision to shoot and the misidentification of harmless objects for guns (Correll, Park, Judd, & Wittenbrink, 2002). But the trial-by-trial manipulation of race raises awareness of the role of race in impacting participants' decisions. Moreover, by the authors' own definitions and arguments, any research that merely correlates responses on implicit measures to behavioral bias cannot on its own be used to support the existence of implicit bias; such a correlation may be obtained for any number of reasons other than the unconscious effects of categories on behavior. Unfortunately, virtually all the research to date arguing for the importance of implicit bias is of this type (see, e.g., responses to the commentaries in Cesario, 2022).

The quality of research supporting potential underlying mechanisms for implicit bias is similarly questionable. Consider the six citations provided by the authors under the "Mechanisms Underlying IB" section. (One wonders why we're talking about mechanisms when we haven't yet established the robustness of the basic effect.) Darley and Gross (1983) has $N = 14$ per cell. Duncan (1976) has $N = 24$ per cell. Gawronski, Geschke, and Banse (2003) has $N = 9$ per cell. Hugenberg and Bodenhausen (2003) has $N = 24$ testing interactions between individual difference and within-subjects manipulations (ps = .04, .02, .02, .04). Kunda and Sherman-Williams (1993) has $N = 8$ per cell. Sagar and Schofield (1980) has $N = 10$ per cell.

Most important, in all these studies and more, the key information is *ambiguous* in nature, a requirement for categories to influence judgments. Given this, in order for implicit bias to impact group disparities, such disparities must be happening in some sizable proportion among the border cases, that gray area in which interpretation can go one way or the other. It is therefore a *requirement* for implicit bias researchers to show that the real-world behavior of interest is ambiguous in nature. As I showed earlier (Cesario, 2022), implicit bias researchers do not do this and as such their claims should not be taken seriously. In response to the second author commenting on this point, I noted:

> A concrete example may help illustrate the irrelevance of experimental findings such as Darley and Gross for explaining real-world disparities. In 2017, there were 13 high schools in the city of Baltimore (student population: over 85% Black, under 5% White) with zero students who tested proficient at grade level in math. In Baltimore's Augusta Fells High School, 50% of students in 2020 had a grade point average (GPA) of 0.13 or lower. This is not the kind of ambiguous, non-diagnostic performance that Ledgerwood et al. would suppose exists in pointing to Darley and Gross as an example of how real- world disparities can be elucidated by experimental social psychology.

Given the authors' own definition, *there is no convincing evidence that implicit bias plays a role in real-world disparities* and researchers must stop claiming otherwise.

Fundamental Questions and Tradeoffs

There is a second, unexplored issue that becomes important once interventions to reduce implicit bias are put on the table, an issue driven by the specific definition of implicit bias proposed in the target article.

It is clear enough how to test for implicit bias in the lab and what such bias demonstrates in a very narrow sense. In a typical experimental design, if targets from different demographic groups are treated differently (without awareness), then evidence for implicit bias is obtained. That is, a target's demographic information can be said to have influenced participants' behaviors toward that target. The question is what this means in a broader sense about human decision-making and the consequences for real-world (rather than simulated in the lab) decision-makers.

If groups are identical on stereotyped traits and behaviors (i.e., if stereotypes are inaccurate), then the influence of categorical information on behavior represents a clear error and efforts to eliminate implicit bias are sound, with benefits almost certainly outweighing any costs. Unfortunately for implicit bias researchers, this is not the case, as groups differ substantially on important characteristics and our most important race- and sex-based stereotypes are overwhelmingly accurate (see Cesario, 2022; Jussim et al., 2016, for discussion and details). Given this, the use of categorical information by decision-makers in generating behavior reflects the appropriate (though not always perfectly accurate) use of prior information in determining behavior, and eliminating the use of such information through implicit bias interventions entails a set of tradeoffs unacknowledged by the authors of the target article.

To illustrate, suppose an implicit bias researcher conducts an experiment in which target sex is manipulated and the dependent variable concerns some violence-preventative behavior (e.g., crossing the street late at night when someone approaches, or calling the police on a possible robbery suspect). The researcher finds that participants treat the target differently (say, being more likely to call the police for male targets) and they do so while unaware of the influence of the target's sex. With the best intentions, the researcher designs an intervention to stop the use of sex in precautionary behaviors and succeeds: people now treat men and

women exactly the same, and the researcher can be satisfied, having changed the world for the better.

The problem is that the sexes differ dramatically in rates of violent crime, and so an unintended consequence for the real decision-makers who were subject to the intervention is that new costs have been imposed on them. Namely, approaching all interactions with men and women as if they have the same potential for violence will increase false negatives, as people will not take precautionary measures when they should have, resulting in more violent victimizations for oneself or others. (None of these costs are experienced by the researcher, safe in her university office.) This also includes, of course, fewer false positives, which is a benefit for those nonviolent males.

To be clear, there may be perfectly good ethical or legal reasons to *not* use categorical information in any such situations. That is not at issue. Similarly, any individual may decide for him or herself that the tradeoffs are "worth it" and stop using categorical information. That also is not at issue. What is at issue is the pretense that implicit bias interventions designed to change behavior entail no tradeoffs or can be imposed *on other people* when the researcher is not the one to bear the relevant costs.

These are not new issues (e.g., Arkes & Tetlock, 2004). The authors sidestep the accuracy and prediction questions but they are fundamental, both to understanding interventions but also to understanding real-world group disparities.

Another Disappointing Research Cycle

The history of implicit bias research provides many cautionary lessons, from poorly defined concepts, to over-extension of the explanatory realm, to political activism far before a reasonable time. It would be wise to learn from these mistakes and revise our approach to research as we move forward. Instead, social psychologists appear poised to repeat this same cycle with their newest grand theory, "systemic bias." As referenced by the authors of the target article, this is yet another vague, broad, extensive framework that is quickly being used by academics in the public sphere to explain every group disparity under consideration.

There is no question that both overt and unspoken discrimination have been directed against various groups over time in the United States and that such discrimination has had harmful effects on the peoples at whom they have been directed. This is not at issue. What is at issue is the broad application of specific historical policies to understanding current disparities, without working through the assumptions necessary for such applications to be sound and without providing detailed data on each assumption.

Consider the housing example used in the target article. The claim is that historical discriminatory housing and lending policies directed at some Black Americans explain current-day disparities. Yet this connection requires a number of steps that are (typically) unstated, for example, that home ownership built wealth and was a major source of intergenerational wealth transfer in ways that would have operated identically for Black and White homeowners during the relevant period (i.e., the counterfactual that those actual Black families would have built wealth in the same ways as White families had such policies not been in place), and that absent such policies other sources of wealth disparities (e.g., spending differences, employee characteristics differences, etc.) either would not have existed or would have been overwhelmed by the counterfactual housing wealth. Moreover, all these would need to be understood in a coherent way alongside the trajectories of other groups, such as Japanese-Americans, who experienced similar policies but who currently have greater wealth than Whites. (Indeed, such broad answers as "systemic racism" or "White supremacy" often fall apart once the groups under consideration are expanded beyond Black and White, or once more detailed distinctions are made, such as between African immigrants and native Black Americans; see, e.g., Sowell, 2009, 2008.) Finally, such assumptions need to be more than just made explicit; they need to be quantified in precise ways and done so for every outcome of interest.

None of these steps are typically taken when social psychologists reference "systemic bias," as illustrated in the target article. Once again, to be clear: This is not to say that such policies did not have negative effects or continue to do so. But if we do not want to be writing these same target articles in another 20 years, it might be a good idea for social psychologists to avoid the same path we have been on for the last 20.

Final Comments on the Authors' Definitions

The authors define implicit bias as *effects* of category information that occur outside awareness. This places implicit bias "outside" the perceiver's head in an important sense, such that it is defined in terms of observed behavioral effects. Such a definition has several advantages, as noted in the target article. However, it is not quite as clean as the authors wish. There is a general tension throughout the article between describing implicit bias as the counterfactual of "that which would not have happened if the category was different" (as implied in all the opening examples) versus "those categorical effects which happen without awareness" (as the formal definition given). The problem with defining implicit bias in this latter sense is that such a definition necessarily entails an individual's *subjective experience*; by definition, then, implicit bias cannot fully be "a behavioral phenomenon" (Gawronski et al., this issue, p. 140) outside the person's head. Furthermore, in any specific case (such as any of the #LivingWhileBlack stories) one cannot know the subjective experience of the actor, so none of those examples clearly qualify as implicit bias.

More important, if implicit bias is merely categorical bias without awareness, one wonders why that specific term is needed at all. The tremendous research output in the 1980s and 1990s on automaticity included lack of awareness as a central variable in this literature. Hence the idea of categories influencing people without their awareness already has a place in the automaticity frameworks of that period (e.g., Bargh, 1989). One can rescue the new definition of implicit

bias by arguing that this earlier automaticity boom was focused primarily on *cognitive processes*, whereas implicit bias is now focused on behavioral effects. But is this really the position to which the authors of the target article want to retreat? After all, regardless of what one might think of the evidential quality of Bargh, Chen, and Burrows (1996)'s classic studies, he was proposing unconscious behavioral effects of categories 25 years ago. So what, then, does the phrase "implicit bias" buy us beyond describing a less nuanced version of the four horsemen of automaticity, applied to behavior only?

There are also two important changes in the authors' definition of implicit bias relative to how such bias has always been understood, at least since the seminal Greenwald and Banaji (1995) article. First, Greenwald and Banaji justified importing the concept of *implicit* from cognitive psychology to social psychology on the grounds that implicit cognition had been thoroughly demonstrated within the cognitive literature (e.g., Jacoby & Witherspoon, 1982). However, the meaning of implicit from that literature referred to unawareness of the *original source* in memory. In Greenwald and Banaji's words:

> Implicit attitudes are introspectively unidentified (or inaccurately identified) *traces of past experience* that mediate favorable or unfavorable feeling, thought, or action toward social objects.

Such a meaning of implicit was not with respect to the effects of categorical information but instead with respect to the origin of those associations that produce categorical effects. After publication of Greenwald and Banaji (1995), the notion of uncontrollability was added to the definition of implicit bias (see, e.g., Nosek, Greenwald, & Banaji, 2007). Perhaps this was done because the main measure developed and used by these authors—the IAT—was patently a measure of controllability and not awareness.

Regardless, the definition proposed in the target article departs meaningfully from past use. Such a departure might be fine, and again there are advantages to the authors' definition. However, if we take the authors seriously and push this definition to its logical conclusion, the three issues raised at the outset do not paint a favorable picture of this research area.

Concluding Comment

When considered seriously, the target article strongly vindicates early and continuing critiques of the implicit bias concept (see, e.g., Arkes & Tetlock, 2004; Blanton & Jaccard, 2008; Blanton, Jaccard, Christie, & Gonzales, 2007, Blanton, Jaccard, Gonzales, & Christie, 2006, Blanton et al., 2009; Cesario, 2022; Corneille & Hütter, 2020; Fiedler, Messner, & Bluemke, 2006; Gawronski, 2019; Machery, 2022; Mitchell, 2017; Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013; Schimmack, 2021). Indeed, serious issues were raised with the conceptualization and measurement of implicit bias *in this very journal* almost two decades ago (Arkes & Tetlock, 2004). As the current target article demonstrates yet again, the strong activism by early implicit bias researchers was far beyond the state of knowledge not only at the time, but even today.

Funding

This work was supported by National Science Foundation Grant No. 1756092.

References

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "would Jesse Jackson 'fail' the implicit association test?" *Psychological Inquiry*, *15*(4), 257–278. doi:10.1207/s15327965pli1504_01

Bargh, J. A. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. *Unintended Thought*, Vol. 1, 3–51.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244.

Blanton, H., & Jaccard, J. (2008). Unconscious racism: A concept in pursuit of a measure. *Annual Review of Sociology*, *34*(1), 277–297. doi:10.1146/annurev.soc.33.040406.131632

Blanton, H., Jaccard, J., Christie, C., & Gonzales, P. M. (2007). Plausible assumptions, questionable assumptions and post hoc rationalizations: Will the real IAT, please stand up? *Journal of Experimental Social Psychology*, *43*(3), 399–409. doi:10.1016/j.jesp.2006.10.019

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the implicit association test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, *42*(2), 192–212. doi:10.1016/j.jesp.2005.07.003

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *The Journal of Applied Psychology*, *94*(3), 567.

Cesario, J. (2022). What can experimental studies of bias tell us about group disparities? *Behavioral and Brain Sciences*, *45*, e66: 1–71.

Corneille, O., & Hütter, M. (2020). Implicit? What do you mean? A comprehensive review of the delusive implicitness construct in attitude research. *Personality and Social Psychology Review*, *24*(3), 212–232. doi:10.1177/1088868320911325

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*(6), 1314–1329. doi:10.1037/0022-3514.83.6.1314

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. doi:10.1037/0022-3514.44.1.20

Duncan, B. L. (1976). Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of blacks. *Journal of Personality and Social Psychology*, *34*(4), 590–598. doi:10.1037/0022-3514.34.4.590

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "i", the "a", and the "t": A logical and psychometric critique of the implicit association test (iat). *European Review of Social Psychology*, *17*(1), 74–147. doi:10.1080/10463280600681248

Gawronski, B. (2019). Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *14*(4), 574–595. doi:10.1177/1745691619826015

Gawronski, B., Geschke, D., & Banse, R. (2003). Implicit bias in impression formation: Associations influence the construal of individuating information. *European Journal of Social Psychology*, *33*(5), 573–589. doi:10.1002/ejsp.166

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, *94*(4), 945–967. doi:10.2307/20439056

Heckman, J. J. (1998). Detecting discrimination. *Journal of Economic Perspectives*, *12*(2), 101–116. doi:10.1257/jep.12.2.101

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, *14*(6), 640–643.

Jacoby, L. L., & Witherspoon, D. (1982). Remembering without awareness. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *36*(2), 300–324. doi:10.1037/h0080638

Jussim, L., Crawford, J. T., Anglin, S. M., Chambers, J. R., Stevens, S. T., & Cohen, F. (2016). Stereotype accuracy: One of the largest and most replicable effects in all of social psychology. In *Handbook of prejudice, stereotyping, and discrimination: Second edition* (pp. 31–63). New York: Taylor and Francis Inc.

Kang, J., & Banaji, M. R. (2006). Fair measures: A behavioral realist revision of affirmative action. *California Law Review*, *94*(4), 1063. doi:10.2307/20439059

Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the construal of individuating information. *Personality and Social Psychology Bulletin*, *19*(1), 90–99. doi:10.1177/0146167293191010

Ledgerwood, A., Pickett, C., Navarro, D., Remedios, J., & Lewis, N. Jr. (2022). The unbearable limitations of solo science: Team science as a path for more rigorous and relevant research. *Behavioral and Brain Sciences*, *45*, e66: 41–45.

Machery, E. (2022). Anomalies in implicit attitudes research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *13*(1), e1569.

Mitchell, G. (2017). Jumping to conclusions: Advocacy and applicationof psychological research. In J. T. Crawford & L. Jussim (Eds.), *The politics of social psychology* (pp. 139–155). Routledge.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). *The implicit association test at age 7: A methodological and conceptual review*. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). Psychology Press.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, *105*(2), 171–192.

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology*, *39*(4), 590–598.

Schimmack, U. (2021). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(2), 396–414. doi:10.1177/1745691619863798

Sowell, T. (2008). *Discrimination and disparities*. New York, NY: Basic Books.

Sowell, T. (2009). *Black rednecks & white liberals*. New York, NY: Encounter Books.