Summary of article being prepared for *Psychological Science in the Public Interest*

Implicit bias remedies: What is known and what is not known

Anthony G. Greenwald, University of Washington
Nilanjana Dasgupta, University of Massachusetts
John F. Dovidio, Yale University
Randall W. Engle, Georgia Tech
Jerry Kang, UCLA (Law)
Corinne Moss-Racusin, Skidmore College
Claude M. Steele, Stanford University
Bethany A. Teachman, University of Virginia

In the article that unveiled *implicit social cognition* as a research area, Greenwald and Banaji (1995) introduced a second phrase, *implicit bias*, which they used just once. It served as a shorthand label for the findings of studies in which indirect measures of intergroup attitudes and stereotypes revealed discriminatory bias. Greenwald and Banaji observed (a) that these indirect measures captured attitudes and stereotypes in ways that were different from what self-report measures revealed, and (b) that this difference possibly would prove important to understanding discriminatory bias. It could not have been predicted that, 25 years later, their once-used phrase (implicit bias) would have gained extremely wide use.

Definition

Kang et al. (2012) offered a useful definition of *implicit bias* as "attitudes or stereotypes that affect our understanding, decision-making, and behavior, without our even realizing it" (p. 1126). This definition works well in multiple contexts because it avoids disagreements that can be raised when definitions attach theoretical meaning to "implicit". The two most frequent such suggestions are (a) that implicit is a synonym of "unconscious" and (b) implicit is a synonym of "automatic". The implicit=unconscious proposal suggests that implicit biases are directed by mental *representations* that are inaccessible to conscious introspection.[1] The implicit=automatic proposal suggests that implicit biases are governed by mental *processes* that operate without awareness. These identifications are problematic, partly because "unconscious" and "automatic"

[1] The interchangeable use of *implicit bias* and *unconscious bias* also characterizes two domains in which relations to types of mental representations have not been of concern. One is in legal scholarship, where "unconscious bias" achieved currency before it was used in psychology. The other is in writing for mass media, where writers may have a preference for one or the other phrase without being concerned about distinguishing their meanings.

do not presently have agreed-upon empirical definitions that can justify interpretation as capturing these representation or process conceptions of "implicit".

The alternative used in this article is to identify "implicit" with the indirect measurement procedures that provided the data reviewed by Greenwald and Banaji. This identification of implicit with *indirectly measured* was later endorsed decisively in a broad review of research on implicit social cognition by Fazio and Olson (2003). Recent discussions of the definition of "implicit" (Greenwald & Banaji, 2017; Greenwald & Lai, 2020) have similarly advocated the interpretation of implicit as "indirectly measured", drawing on previous strong analyses of the challenge of identifying any mental process as purely unconscious or automatic (Jacoby, 1991; Reingold & Merikle, 1988).

Public understanding

Media descriptions of scientific work on implicit bias in the past 20 years have created public awareness and concern about a form of discrimination that can occur without appearance of perpetrators' intent to discriminate. Many media presentations describe implicit bias as a widely acknowledged cause of disparities in outcomes or benefits associated with differences in race, gender identity, ethnicity, age, disability, socioeconomic status, sexual orientation, and more. Accompanying the public education value of these media presentations, there are some unfortunate side effects. Wide awareness of implicit bias has prompted entrepreneurs to offer training programs that are advertised as providing remedies for undesired discriminatory consequences of implicit biases. Such programs often have three components: (a) defining implicit bias as a source of unintended discrimination, (b) describing the pervasiveness of implicit biases, and (c) offering remedies to combat effects of implicit biases on judgment and decision making. Components (a) and (b) often draw relatively accurately on scientific understanding, but (c) is often presented without foundation in relevant research and with no effort to demonstrate effectiveness of proffered remedies. One purpose of this article is to describe scientific work on implicit bias in ways that can usefully inform those who offer, those who purchase, and those who participate in such educational and training efforts.

Trainers' advocacy of empirically unsupported remedies for unintended discrimination has the further downside of creating an unjustified appearance that the organization hosting the training is operating in bias-free fashion. Kaiser et al. (2013) concluded that the "illusory sense of fairness derived from the mere presence of diversity structures causes high-status [e.g., racial majority] group members to legitimize the status quo by becoming less sensitive to discrimination targeted at underrepresented groups and reacting more harshly toward underrepresented group members who claim discrimination" (p. 504). Studies of effectiveness of diversity programs have not only failed to find evidence that they improve bias-free hiring, but have often found that the training is counterproductive. Kalev et al. (2006) reviewed three classes of diversity-improving strategies (diversity training, networking with mentoring, and establishing organizational responsibility) in 708 private-sector companies, using data from those companies' annual EEO-1 hiring reports (to the U.S. Equal Employment Opportunity Commission). Kalev et al. concluded that "Efforts to moderate managerial bias through diversity

training and diversity evaluations are least effective at increasing the share of white women, black women, and black men in management" (p. 589). They found that "establishing organizational responsibility" was the only one of the three categories of methods for which they could find evidence of effectiveness in increasing hiring diversity.

Limitations of most currently offered training programs have two roots in present scientific understanding. One is the incompleteness of scientific understanding of implicit bias (the 'what is not known' of this article's title). The second is that important conclusions from research on implicit bias (the 'what is known' of the title) have yet to penetrate to many of those who are offering remedial programs. This article's identifications of what is known and what is not known are intended to enable wider understanding of the inadequacy of some suggested remedies, which may prompt more effective use of both private and public resources available for the purpose of reducing inequities.

MISUNDERSTANDINGS OF IMPLICIT BIAS

Before taking up what is known scientifically about implicit bias, we should dispel misunderstandings of implicit bias that have gained footholds, either from media presentations that were insufficiently science-informed or from scientific presentations that are now known to have been premature**.**

Misunderstanding: The IAT and other indirect measures assess 'implicit prejudice' or 'implicit racism'

Within a few years after the first publication of the IAT, the measure's creators and its most active developers stopped using the words 'prejudice' or 'racism' in published descriptions of what the IAT measured. There were three reasons for this change: First, an accumulation of findings had by then made clear the divergence between what was revealed by self-report measures and what was revealed by parallel IAT measures. Second, nothing about the IAT's procedure should prompt subjects, while their classification latencies were being recorded, to have in mind the hostility or antipathy that is central to most published definitions of prejudice (Greenwald & Pettigrew, 2014, p. 684). Third, an IAT score indicating preference for racial White relative to Black can be obtained by someone who likes both racial groups, but likes Whites more. In contrast with IAT measures, self-report measures of racial attitudes typically oblige subjects to actively contemplate hostile or disparaging sentiments about outgroups while reporting their agreement or disagreement with those sentiments.

Misunderstanding: Implicit and explicit biases are uncorrelated (empirical data show that parallel implicit and explicit measures of biases are very generally positively correlated)

Not knowing more than that important distinctions between implicit and explicit measures of race attitudes had been empirically established, many assumed that these measures were uncorrelated with one another and, more generally, that parallel implicit and explicit measures of the same attitude were generally uncorrelated. As the "What is Known" section's review of established findings will indicate, parallel implicit and explicit measures are almost uniformly positively correlated, with these correlations sometimes being large. Nevertheless, the

distinctions between what is measured by the two types of procedures are very substantial, which will also be described.

Misunderstanding: Implicit biases are malleable (empirical research has yet to establish practically useful methods of durably weakening long-established implicit biases)

The Oxford English Dictionary defines 'malleable' as "capable of being hammered or pressed out of shape *without* a tendency to return to the original shape". The conclusion that changes in IAT measures following brief interventions indicate *malleability* of implicitly measured attitudes or stereotypes therefore implies that the observed changes are durable. Between the dates of a literature review by Blair (2002) and a set of multiple experimental studies by Lai et al. (2014), empirical findings from intervention studies were widely interpreted as supporting the "malleable" conclusion. Only with a new series of studies by Lai et al. (2016) did it become clear that almost all published intervention studies had used relatively brief single sessions, with posttests on IAT measures collected typically within about 15 minutes of the intervention. Lai et al. (2016) tested the same interventions (i.e., those previously found to be effective with brief delays) additionally with delays ranging from several hours to several days, finding that none was effective with the longer durations, thereby undermining the prior belief that implicit attitudes were easily and durably modifiable. This history is described more fully in the "What is Known" section of this article.

Misunderstanding: Attempts at implicit-bias remediation are often use empirically untested (or insufficiently tested) strategies.

Under-tested interventions include instructing audiences to form and remember intentions to avoid bias or advising them to go slowly in making decisions that might be biased (e.g., pausing to think, meditating). The very small amount of available evidence concerning these strategies is considered in this article's "What is Known" section, alongside evidence concerning strategies that can be more confidently suggested.

References

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review*, *6*, 242–261

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and uses. *Annual Review of Psychology*, 54, 297–327.

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

Greenwald, A. G., & Banaji, M. R. (2017). The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist, 72*, 861–871.

Greenwald, A. G., & Lai, C. K. (2020[in press]). Implicit social cognition. *Annual Review of Psychology*.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*, 1464–1480.

Greenwald, A. G, Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216.

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist, 69*, 669–684.

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

Jacoby, L. L. (1991). A process-dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language, 30*, 513–541.

Kaiser, C. R., Major, B., Jurcevic, I., Dover, T. L., Brady, L. M., & Shapiro, J. R. (2013). Presumed fair: Ironic effects of organizational diversity structures. *Journal of Personality and Social Psychology, 104*(3), 504-519.

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review, 71*, 589–617.

Kang, J., Bennett, M. W., Carbado, D. W., Casey, P., Dasgupta, N., Faigman, D. L., Godsil, R. D., Greenwald, A. G. , Levinson, J. D., & Mnookin, J. L. (2012). Implicit bias in the courtroom. *UCLA Law Review, 59*, 1124–1186.

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

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

Reingold, E. M., & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics*, *44*, 563-575.