AN IMPLICIT BIAS PRIMER

Gregory Mitchell[*]

[*] Joseph Weintraub – Bank of America Distinguished Professor of Law, University of Virginia School of Law, 580 Massie Road, Charlottesville, VA 22903, 434-243-4088, greg.mitchell@virginia.edu. An earlier version of this paper was presented at the Practising Law Institute's program titled "Fact or Fiction – Implicit Bias in the Workplace and Courtroom."

AN IMPLICIT BIAS PRIMER

Gregory Mitchell

The phenomenon of implicit bias is much discussed but little understood. This article answers basic conceptual and empirical questions about implicit bias, including what it is, how it is measured, what effects it may have on behavior, and whether it can be changed. Consensus now exists among implicit bias researchers that current measures of implicit bias cannot reliably identify who will or will not discriminate in any given situation and that programs aimed at changing implicit bias produce very limited effects. Despite hopes that implicit bias research would lead to new and better understandings of how and why discrimination occurs, the empirical reality is that implicit bias research has not yet improved our ability to predict and prevent discrimination.

INTRODUCTION

THE concept of implicit bias looms large in contemporary discussions of the origins of discrimination against women and minorities.1 In both the popular press and law reviews, the view that implicit bias causes many acts of discrimination has become common.2 New York Times columnist Nicolas Kristof recently wrote, for instance, that "the evidence is overwhelming that unconscious bias remains widespread in ways that systematically benefits whites and men."3 Federal District Judge Mark Bennett extends this view to the courtroom:

> [W]e unconsciously act on implicit biases even though we abhor them when they come to our attention. Implicit biases cause subtle actions . . . [b]ut they are also powerful and pervasive enough to affect decisions about whom we employ, whom we leave on juries, and whom

[1] "Perhaps no new concept from the world of academic psychology has taken hold of the public imagination more quickly and profoundly in the 21st century than implicit bias — that is, forms of bias which operate beyond the conscious awareness of individuals." Jesse Singal, Psychology's Favorite Tool for Measuring Racism Isn't Up to the Job, NY Mag (Jan. 11, 2017, 12:18PM), http://nymag.com/scienceofus/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html.

[2] A search of Westlaw's Secondary Sources database for sources discussing implicit bias ("implicit bias") produces more than 2,300 hits, and a search for sources discussing implicit bias and its relation to behavior ("implicit bias" /s behavior) produces more than 350 hits.

[3] Nicolas Kristof, Opinion, Straight Talk for White Men, N.Y. Times, Feb. 21, 2015, http://www.nytimes.com/2015/02/22/opinion/sunday/nicholas-kristof-straight-talk-for-white-men.html.

> we believe. Jurors, lawyers, and judges do not leave behind their implicit biases when they walk through the courthouse doors.[4]

In addition to injustices in the courtroom, police misconduct, health disparities, housing disparities, employment disparities, educational disparities, and everyday slights and insults have all been attributed to implicit bias.[5] Implicit bias has become an all-purpose explanation for societal problems and poor intergroup relations. Indeed, while recent events demonstrate that old-fashioned racism and anti-Semitism are far from dead,[6] some hold the view that implicit bias "plays a larger role in

[4] Mark W. Bennett, Unraveling the Gordian Knot of Implicit Bias in Jury Selection: The Problems of Judge-Dominated Voir Dire, the Failed Promise of Batson, and Proposed Solutions, 4 Harv. L. & Pol'y Rev. 149, 150 (2010). A few recent examples of the many similar statements found in the law reviews include: Michele Goodwin & Erwin Chemerinsky, No Immunity: Race, Class, and the Constitutional Implications of Inoculation, 129 Harv. L. Rev. 956, 987 (2016) ("Implicit biases shape unconscious attitudes, beliefs, and actions according to status markers such as ethnicity, religion, sex, class, and race." (footnote omitted)); Alex B. Long, Employment Discrimination in the Legal Profession: A Question of Ethics?, 2016 U. Ill. L. Rev. 445, 449–50 ("employment discrimination is often a problem of implicit biases and institutionalized obstacles to equal employment opportunity"); Russell G. Pearce, Eli Wald & Swethaa S. Ballakrishnen, Difference Blindness vs. Bias Awareness: Why Law Firms with the Best of Intentions Have Failed to Create Diverse Partnerships, 83 Fordham L. Rev. 2407, 2413 (2015) ("Lawyers bring to their work their implicit biases that are embedded in the dominant power and prestige of identity groups in society. To the extent that white men are the dominant group in society, leaders of law firms will bring biases in their favor into the workplace." (footnotes omitted)); Russell K. Robinson, Unequal Protection, 68 Stan. L. Rev. 151, 157 (2016) ("[A] plausible reading of Obergefell provides an opening for courts to draw on the science of implicit bias to provide a more reliable and objective anchor for understanding bias not just in sexual orientation cases, but in equal protection cases more generally. Although this is merely implied in Obergefell, I urge future courts and scholars to build on this opportunity.").

[5] See, e.g., Mahzarin R. Banaji & Anthony G. Greenwald, Blindspot: Hidden Biases of Good People 193–200 (2013) (attributing disparities in housing, hiring, health care, and criminal justice in part to implicit bias); Janel A. George, Stereotype and School Pushout: Race, Gender, and Discipline Disparities, 68 Ark. L. Rev. 101, 109–13 (2015) (attributing group differences in school discipline to implicit bias).

[6] See, e.g., Hawes Spencer, A Far-Right Gathering Bursts Into Brawls, N.Y. Times (Aug. 13, 2017), https://www.nytimes.com/2017/08/13/us/charlottesville-protests-unite-the-right.html?_r=0.

discriminatory behavior today than the explicit biases on full display in the 1960s when Title VII of the 1964 Civil Rights Act was passed."[7]

Unfortunately, many popularizations of the implicit bias concept, both in law reviews and mainstream media, rely on statements made during the first generation of implicit bias research, when there was great optimism about the power of measures of implicit bias to identify persons who are more and less likely to engage in acts of discrimination—but little data at that time to support such optimism. The second generation of implicit bias research has produced decidedly less optimistic views about the predictive and explanatory power of implicit bias measures. This second generation of research has also complicated our understanding of the nature of implicit bias and how it may relate to explicit bias. Using the latest research on implicit bias, including large-scale meta-analyses of the full body of research, this primer summarizes our current understanding of what implicit bias is, how to measure it, how it may relate to behavior, and whether it can be altered.

I. WHAT IS IMPLICIT BIAS, AND HOW IS IT MEASURED?

Colloquially, bias refers to a tendency to favor or disfavor something, whether it be one's right hand when dribbling a basketball or a particular political party when voting. Within psychology, bias has a similar meaning: Bias refers to the tendency to react differently to stimuli based on particular characteristics of the stimuli. For instance, if one tends to react more positively to White persons than Black persons, then one is *biased* in favor of Whites and against Blacks. If this bias operates at the level of affective reactions (i.e., good or bad feelings), then we say that one holds *prejudicial attitudes* toward Whites and Blacks. If this bias operates at the level of personal or behavioral characteristics assigned to a group, then we say that one holds *stereotypes* about Whites and Blacks (e.g., perhaps one believes Blacks are generally more athletic than Whites).[8]

[7] Sahar F. Aziz, Coercing Assimilation: The Case of Muslim Women of Color, 18 J. Gender Race & Just. 389, 391 (2016) (footnote omitted); accord Linda Hamilton Krieger, The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity, 47 Stan. L. Rev. 1161, 1164 (1995) ("I argue that the way in which Title VII jurisprudence constructs discrimination, while sufficient to address the deliberate discrimination prevalent in an earlier age, is inadequate to address the subtle, often unconscious forms of bias that Title VII was also intended to remedy. These subtle forms of bias, I suggest, represent today's most prevalent type of discrimination." (footnote omitted)); cf. L. Song Richardson, Systemic Triage: Implicit Racial Bias in the Courtroom, 126 Yale L.J. 862, 892 (2017) ("this new racism is in some ways more dangerous and pernicious than racial bigotry because it is ephemeral and difficult to eradicate").

[8] See John Duckitt, Prejudice and Intergroup Hostility, in Oxford Handbook of Political Psychology 559, 559 (David O. Sears et al. eds., 2003) ("So-

Bias can be assessed in one of two ways: using direct or indirect measures.[9] With a direct measure, the researcher asks someone to report orally or in writing how one feels about a particular group or what traits one associates with a particular group, and often the questions seek comparative responses (i.e., relative warmth toward Whites vs. Blacks, or the rating of both groups on traits such as athleticism). If a bias is evident on these direct measures, then we say that the person is explicitly biased toward a group (e.g., if the respondent rates one group as more athletic than the other, then we would say that the respondent holds an explicit bias toward these groups that takes the form of an explicit athleticism stereotype).

For some types of bias, researchers worry that social norms may discourage honest reporting on direct measures of bias.[10] For instance, in some settings, people may be reluctant to report honestly on their political views, and persons who harbor a bias against a particular group may worry about the social consequences of admitting that bias. Therefore, direct measures of bias may sometimes fail to obtain accurate data on people's biases because people are hiding their biases.

cial psychologists have distinguished three distinct components of prejudice, or ways in which negative intergroup attitudes can be expressed. These are negative stereotypes (cognitive component), negative feelings (affective component), and negative behavioral inclinations (behavioral component) toward outgroups."). Bias involves systematic deviation from some benchmark, and that benchmark need not be derived from an assumption of equal relative treatment. For instance, many studies of judgment and decision-making define bias in terms of performance relative to the predictions of rational choice theory. See Gregory Mitchell, Why Law and Economics' Perfect Rationality Should Not Be Traded for Behavioral Law and Economics' Equal Incompetence, 91 Geo. L.J. 67, 80-81 (2002). When intergroup bias is defined in relative terms, a bias may be found even when both groups elicit positive or negative reactions (e.g., one may feel positive about both Whites and Blacks, but one may have a more positive reaction to one group than the other).

[9] See Gregory Mitchell & Philip E. Tetlock, Implicit Attitude Measures, in Emerging Trends in the Social and Behavioral Sciences (Stephen Michael Kosslyn, ed., 2015), http://onlinelibrary.wiley.com/book/10.1002/9781118900772?campaign=rdtwol-42073.671909722&elq_mid=3751&elq_cid=1424663 (discussing indirect, or implicit, measures in detail).

[10] See id.; see also Shanto Iyengar et al., Understanding Explicit and Implicit Attitudes: A Comparison of Racial Group and Candidate Preferences in the 2008 Election 3 (2008), https://pcl.stanford.edu/research/2010/iyengar-understanding.pdf ("To some extent, the sharp decline in self-reported racial prejudice may represent an artifact of survey research rather than meaningful attitude change. In the social (and sometimes interpersonal) setting of an opinion survey, whites may be motivated to conform to widely shared egalitarian norms and respond in a manner that suggests the absence of racial bias" (citation omitted)). See generally Roger Tourangeau & Ting Yan, Sensitive Questions in Surveys, 133 Psychol. Bull. 859 (2007).

Another worry with direct measures is that some biases may be difficult to access through introspection. Direct measures of bias assume that people can examine their own minds and determine what biases lurk there. However, bias may be the product of mental processes and structures that cannot be easily accessed through conscious thought.[11] Therefore, direct measures of bias may also produce inaccurate or incomplete data because people are unaware of their biases.

These concerns about candor and introspective access gave rise to the use of indirect measures of bias. With indirect measures, a researcher asks someone to react to stimuli on some task that does not require verbal reporting of attitudes or beliefs about the stimuli, and then the researcher infers from the pattern of responses to the stimuli whether a bias may exist. For instance, in the most popular indirect measure of bias, the Implicit Association Test ("IAT"), respondents are asked to categorize stimuli that flash onto a computer screen as fast as they can.[12] In the racial attitudes IAT, respondents are presented with four types of stimuli: words representing positive feelings (e.g., good), words representing negative feelings (e.g., bad), pictures of faces of Black persons, and pictures of faces of White persons. In one portion of the IAT, if a positive word or White face appears, then the "E" key on a computer keyboard is to be pressed as fast as possible and if a negative word or Black face appears the "I" key is to be pressed as fast as possible. On a separate portion of the IAT, if a positive word or Black face appears the "E" key is to be pressed as fast as possible and if a negative word or White face appears the "I" key is to be pressed as fast as possible. If the respondent responds relatively more quickly on the first portion of the test than the second portion of the test, then the researcher infers that the test-taker holds an implicit positive bias toward Whites and an implicit negative bias toward Blacks. The idea is that stimuli that are affectively compatible should be easier to categorize than stimuli that are affectively incompatible, presumably because one need not resolve any tension between two compatible items when deciding how to categorize them.

In sum, implicit bias is the bias that is "implied" by an individual's performance on an indirect measure of bias. Although the IAT is the most popular indirect measure of bias, several other indirect measures of

[11] See, e.g., Anthony G. Greenwald et al., Implicit Race Attitudes Predicted Vote in the 2008 U.S. Presidential Election, 9 Analyses Soc. Issues & Pub. Pol'y 241, 242 (2009) ("Unlike explicit attitude measures, which assume awareness of the attitudes being assessed, implicit attitude measures do not require awareness and can capture attitudes that may be introspectively inaccessible.").

[12] The IAT was introduced in 1998. See Anthony G. Greenwald et al., Measuring Individual Differences in Implicit Cognition: The Implicit Association Test, 74 J. Personality & Soc. Psychol. 1464, 1465 (1998) (providing a schematic description of the IAT). Demonstration tests can be taken online at https://implicit.harvard.edu/implicit/.

bias exist.[13] Explicit bias is the bias observed when an individual verbally expresses a negative or positive attitude toward members of a particular group when directly asked how the individual feels about that group, or when an individual explicitly endorses a positive or negative stereotype with respect to a particular group.

II. WHAT CAUSES THE RESPONSE PATTERNS FROM WHICH IMPLICIT BIAS IS INFERRED?

Implicit bias is commonly portrayed as reflecting associative networks that exist within the brain, in which repeated pairings of stimuli supposedly lead to stronger associations between two items.[14] Under this view, watching lots of NBA games in which Black players are more common than White players might give rise to a stronger association between "Black person" and "athleticism." From the associative network view, any time a stimulus is encountered, then its area within the associative network becomes activated, with other connected nodes activated as well (i.e., there is spreading activation within the network).[15] Thus, encountering a Black person should activate various affective and semantic associations within the network, including the athleticism node. Note, however, that these associative networks may involve a complex mix of positive, neutral, and negative associations. The particular context in which a stimulus is perceived, and thus the other nodes likely to

[13] See Bernd Wittenbrink & Norbert Schwarz, Introduction to Implicit Measures of Attitudes 1, 4–8 (Bernd Wittenbrink & Norbert Schwarz eds., 2007) (overview of various indirect measures). The IAT and other indirect measures of bias have been used to study a wide array of possible biases, from racial, ethnic, and gender biases to biases against religious groups, political groups, and even cats versus dogs. For examples of the many variations of the IAT that have been developed, see the appendix to Brian A. Nosek, Moderators of the Relationship Between Implicit and Explicit Evaluation, 134 J. Experimental Psychol. 565 (2005).

[14] See Eric Mandelbaum, Attitude, Inference, Association: On the Propositional Structure of Implicit Bias, 50 Noûs 629, 630 (2016) ("A far less trivial, but no less widely accepted claim is that implicit bias is, in an important sense, caused by some associative process/associative structure. The associative process/structure is generally assumed but not discussed in any depth. Insofar as its meaning is analyzed, it is usually glossed as some type of evaluative association, such as an association between a valence (e.g., negative affect) and a concept (e.g., BLACK MALE)." (footnote omitted)).

[15] See, e.g., Perry Hinton, Implicit Stereotypes and the Predictive Brain: Cognition and Culture in "Biased" Person Perception, 3 Palgrave Comm. 1, 2 (2017) ("Concepts in semantic memory are assumed to be linked together in terms of an associative network, with associated concepts having stronger links, or are closer together, than unrelated concepts. . . . Activation of one concept (such as reading the word 'doctor') spreads to associated concepts in the network (such as 'nurse') making them more easily accessible during the activation period." (citations omitted)).

be simultaneously activated, will have an important influence on the response given to the stimuli.[16] For instance, when researchers portrayed the faces in the racial attitudes IAT as churchgoers, the usual observed bias toward Blacks was no longer observed and respondents showed no measured bias for Blacks or Whites.[17] Therefore, the many, sometimes competing, associations that we have to various groups, along with the possible influence of fleeting situational cues in activating associations, make it very difficult to predict when any particular bias will be activated.

Notwithstanding theorizing that implicit bias is the product of associations learned over time, recent research casts doubt on the associative model of implicit bias. In particular, there is now evidence that propositions can account for many of the response patterns observed on indirect measures of bias.[18] Whereas the mental association construct represents the frequency with which two items have been experienced together and thus the likelihood that experiencing one item will bring to mind the other item (e.g., observing a match may bring to mind fire) regardless of their logical or functional relationship, a mental proposition reflects the

[16] Bertram Gawronski, Skylar M. Bannon & Galen V. Bodenhausen, The Associative-Propositional Duality in the Representation, Formation, and Expression of Attitudes, in Reflective and Impulsive Determinants of Human Behavior 103, 110–11 (Roland Deutsch et al. eds., 2017) ("An important aspect of spreading activation is that it is not an all-or-none process, such that encountering a given object would activate each and every concept that is mentally linked with that object in memory. Instead, activation typically spreads only to a limited subset of associated concepts. Which subset is activated in response to an object is assumed to be constrained by the overall configuration of input stimuli, including both the target object and the context in which it is encountered. For example, encountering an African American man in a jazz bar may activate the stereotypical attribute musical, whereas the same African American man may activate the stereotypical attribute criminal if he is encountered in a dark alley." (citation omitted)).

[17] Jamie Barden, William W. Maddux, Richard E. Petty & Marilyn B. Brewer, Contextual Moderation of Racial Bias: The Impact of Social Roles on Controlled and Automatically Activated Attitudes, 87 J. Personality Soc. Psychol. 5, 16 (2004) ("the prisoner role produced in-group bias, the churchgoer role produced equal evaluation, and the factory worker role produced out-group bias").

[18] See Jan De Houwer, A Propositional Model of Implicit Evaluation, 8 Soc. & Personality Psychol. Compass 342, 345 (2014) ("A propositional model of implicit evaluation postulates that a stimulus can evoke an evaluative response automatically only after a proposition about the evaluative properties of the stimulus has been formed or activated automatically."); id. at 342 ("In line with such a propositional model of implicit evaluation, evidence suggests that implicit evaluation (a) can be based on instructions and inferences, (b) is sensitive to information about how stimuli are related and (c) can reflect several propositions that differ only with regard to how stimuli are related.").

relation between two items, such as "Whites are privileged," "Blacks have suffered," or "matches cause fire."

Support for the propositional model of implicit bias comes from a variety of studies.[19] For instance, several studies find that exposing people to propositional information alone affects responses on IATs, which should not happen if the responses are based on learned associations.[20] These studies examine reactions toward novel stimuli for which no preexisting associations could have been learned; only the propositional information to which test-takers have been exposed could explain any differential responses to the stimuli that were observed.[21] Other studies have found that attempting to change implicit biases through conditioning (i.e., repeated pairings of stimuli to alter existing associations) were unsuccessful, contrary to the prediction of an associative model.[22] But exposure to propositional information (i.e., statements that undercut the supposed relation between two items that could have given rise to the bias) did change reactions to the stimuli.

Evidence in favor of the propositional model of implicit bias has at least two important implications. First, it complicates interpretations of data on indirect measures of bias. If the IAT or other indirect measures of bias are being affected by propositions that reflect some relationship among stimuli, rather than just associations among items, then it will be important to understand the nature of the underlying propositional relations. For instance, if a majority of respondents more quickly categorize Black faces when they are paired with negative words than positive words (as is often observed), does that pattern reflect a stronger negative association with the category Blacks or does it reflect some proposition such as "Blacks have it bad in American society" or "I feel bad for

[19] See Mandelbaum, supra note 14, at 637–48 (summarizing studies inconsistent with the associative network view of implicit bias).

[20] De Houwer helpfully contrasts propositions and associations:

> Propositions are statements about the world. For instance, the proposition "I am good" is a statement about the properties of oneself. Importantly, as statements about world, propositions specify the nature of the relation between concepts. For instance, the propositions "I am good" and "I want to be good" both involve the concepts "I" and "good" but differ in the way in which these concepts are said to be related. Hence, a vital difference between propositions and associations is that propositions contain relational information, that is, information about how concepts are related. It is often also assumed that propositions about events can be formed not only on the basis of the repeated experience of those events but also as the result of a single instruction or inference concerning those events.

De Houwer, supra note 18, at 344.

[21] Mandelbaum, supra note 14, at 643–46.

[22] Id. at 638–39.

Blacks"?[23] Under this account, to understand the nature of the implicit bias, one must understand the nature of the propositions that underlie the observed bias (and most indirect measures of bias cannot adjudicate between the associative and propositional account without special steps being taken to attempt to do this).

Second, the propositional account of implicit bias affects views about how implicit biases may be created and changed. Rather than emphasizing the need for more positive images of minorities in the workplace or the media, the focus shifts to presenting people with more detailed or different propositional information that they will accept and take on as part of their cognitive architecture for thinking about groups.

Another complication in understanding the origins of implicit bias involves the distinction between the mental structures that may give rise to implicit bias—whether those involve associations, propositions, or a mix of both[24]—and the influence of these structures on responses to

[23] Support for the latter interpretation has been found in studies that examine whether empathy for the plight of Blacks can explain the common pattern observed on the race IAT in which Black faces are easier to categorize when paired with negative words. See Michael R. Andreychik & Michael J. Gill, Do Negative Implicit Associations Indicate Negative Attitudes? Social Explanations Moderate Whether Ostensible "Negative" Associations Are Prejudice-Based or Empathy-Based, 48 J. Experimental Soc. Psychol. 1082, 1092 (2012) ("The present work challenges the notion that implicit negative associations necessarily reflect negative or prejudiced attitudes. Drawing on work linking external explanations to empathic responses, four studies provided evidence consistent with the argument that whereas implicit 'negative' associations are prejudice-based among individuals who reject external explanations, such implicit 'negative' associations are empathy-based among individuals who embrace external explanations."); Eric Luis Uhlmann, Victoria L. Brescoll & Elizabeth Levy Paluck, Are Members of Low Status Groups Perceived as Bad, or Badly Off? Egalitarian Negative Associations and Automatic Prejudice, 42 J. Experimental Soc. Psychol. 491, 498 (2006) ("The present research suggests that automatic associations have roots in both prejudiced (e.g., 'African Americans are stupid, lazy, and violent') and egalitarian (e.g., 'African Americans are oppressed, mistreated, and victimized') sentiments. Egalitarian negative associations may contribute to prejudiced reactions, as suggested by recent work on the predictive validity of association based measures.").

[24] As the research presently stands, the evidence is mixed as to whether the associative model, propositional model, or a mix of the two best explains the observed implicit bias data. See Gawronski et al., supra note 16, at 115 ("In cases where theoretical disagreements involve genuine empirical issues, there is substantial evidence supporting the joint operation of associative and propositional processes (e.g., the involvement of associative and propositional processes in the formation of evaluative representations). As we noted, single-process propositional theories may be revised to account for these fmdings in a post-hoc fashion, but the dual process approach seems superior because it predicts them a priori without requiring the addition of ad hoc assumptions."); Mandelbaum, supra note 14, at 648 ("It's time to stop assuming that since a process is uncon-

stimuli. Just because one may mentally possess a positive or negative association or a proposition relevant to some social group, that does not mean that structure will necessarily determine how one responds to members of that group.

Within the associative model, in which spreading activation across an associative network automatically occurs when encountering a stimulus without any conscious effort on the part of the perceiver, the nature and extent of that activation will depend on the context, including what other stimuli are perceived and what other memories may come online at the time. The spreading activation model is not, however, the only model of the psychological processes engaged when perceiving members of a social group. Prominent alternatives involve a combination of associative and propositional processes or metacognitive processes, in which association-based responses to a stimulus are constrained by propositions or metacognitions (higher-level cognitions). For example, the propositional information, if accessed, may override any automatic associations activated or cause them to be interpreted in a way that is consistent with a proposition. Although existing empirical evidence suggests that propositional and metacognitive processes serve as important constraints on associative reactions, presently there is no single dominant model of the psychological processes that control the expression of implicit bias.[25]

scious, or automatic, or fast, then it must be associative. Unconscious, automatic, and fast processes can be logical and can operate over propositional structures.").

[25] See, e.g., Gawronski et al., supra note 16, at 105 ("Associations can be understood as dormant links between nodes that constrain the spread of activation within associative networks. Activated patterns of associations, in turn, are assumed to provide the basis for momentarily constructed propositions about states of affairs. From this perspective, any proposition is based on patterns of activated associations; there is no association-independent storage of propositional statements in a different part of long-term memory."); Mandelbaum, supra note 14, at 646 ("Although we've only been discussing the relation between associations and implicit attitudes, it's worth noting that the pure associative story has been long dead in other parts of cognitive science. Not many psycholinguists take associative structure to be the only type of representational structure. This is because one really can't do psycholinguistics (never mind generative semantics or syntax) without, at a minimum, structures that take truth-values, and because associations aren't truth-apt, they cannot serve that role."); Richard E. Petty, A Metacognitive Model of Attitudes, 33 J. Consumer Res. 22, 22 (2006) ("In our metacognitive model, attitudes are conceived as object-evaluation links but with the addition of validation tags. . . . The model recognizes that not only do people have quick evaluative associations to a wide variety of attitude objects, but they also may consider whether these evaluative associations are valid reflections of their own personal assessment of the object. This validity checking . . . need not occur online but with rehearsal can stem from a prestored tag." (citations omitted)); Jeffrey W. Sherman, Controlled In-

A final complication in understanding what exactly gives rise to (the pattern of responses interpreted as) implicit bias is the fact that indirect measures of bias are noisy.[26] By their very nature, indirect measures engage respondents on one task and then use performance on that task to infer bias, but ability to perform the instructed task will always affect the pattern of responses observed and make interpretation of that pattern difficult. Because the IAT is cognitively demanding, the age and general cognitive processing speed of the respondents have been found to affect the pattern of results observed.[27] Likewise, because the IAT becomes easier with practice, the pattern of responding observed on IATs tends to look less biased with practice (i.e., there are fewer differences in reactions to the stimuli drawn from different groups).[28] Generally, the IAT and other indirect measures of bias are subject to noise from the testing

fluences on Implicit Measures: Confronting the Myth of Process-purity and Taming the Cognitive Monster, in Attitudes: Insights from the New Wave of Implicit Measures 391, 391 (Richard E. Petty et al. eds., 2009) ("Though implicit measures are often portrayed as process-pure measures of automatic attitudes, instead, they reflect the joint contributions of automatic and controlled processes."). For discussion of dual-process models more generally, see Gregory Mitchell, Second Thoughts, 40 McGeorge L. Rev. 687 (2009).

[26] See Hart Blanton & James Jaccard, Unconscious Racism: A Concept in Pursuit of a Measure, 34 Ann. Rev. Soc. 277, 287–90 (2008) (discussing sources of noise that confound interpretations of IAT scores).

[27] *See, e.g.*, Mary Lee Hummert et al., *Using the Implicit Association Test to Measure Age Differences in Implicit Social Cognitions*, 17 PSYCHOL. & AGING 482, 492 (2002) ("These two studies show that the IAT may be effectively used to test theoretical explanations about age group differences in implicit attitudes, but they also illustrate that age-related differences in processing speed must be considered in analyzing and interpreting IAT results."); Brian A. Nosek, Anthony G. Greenwald & Mahzarin R. Banaji, *The Implicit Association Test at Age 7: A Methodological and Conceptual Review*, *in* AUTOMATIC PROCESSES IN SOCIAL THINKING AND BEHAVIOR 265, 272 (John A. Bargh ed., 2007) ("Perhaps related to effects of variations in cognitive fluency, older subjects tend to show larger IAT effects than younger subjects, especially when the original scoring algorithm is used. The improved scoring algorithm . . . reduces this relationship between age and IAT scores." (citations omitted)); Cornelia Wrzus, Boris Egloff & Michaela Riediger, *Using Implicit Association Tests in Age-Heterogeneous Samples: The Importance of Cognitive Abilities and Quad Model Processes*, 32 PSYCHOL. & AGING 432, 440 (2017) ("The current study demonstrated that individual differences in cognitive abilities and quad process parameters contribute to IAT scores in age-heterogeneous samples, however not significantly differently with age.").

[28] See Nosek et al., supra note 27, at 273 ("Effect magnitudes with the IAT tend to decline with repeated administrations."). For instance, in the publicly-archived data from Project Implicit's race IAT (https://osf.io/52qxl/), the average IAT D-score declines from .33 on the first administration of the test to .29 on the second administration, to .26 on the third administration, and down to .21 after seven or more administrations of the test.

environment and features of the test-taker that have nothing to do with the associations or propositions that test-takers may hold with respect to members of a particular group.[29] As a result, indirect measures of bias have been found to have low to moderate test-retest reliability[30] (e.g., the score one receives on an IAT on Monday may differ considerably from the score one receives on the same test on Friday), and scores from different indirect bias measures often show little agreement (e.g., one who supposedly shows high implicit bias on the IAT may show low implicit bias on another indirect measure of bias).[31]

III. IS IMPLICIT BIAS SYNONYMOUS WITH UNCONSCIOUS BIAS?

Another common misconception found in popular discussion of implicit bias is that implicit bias is synonymous with unconscious bias.[32]

[29] See, e.g., Blanton & Jaccard, supra note 26, at 288 ("It is well known that latencies on trials within a block can vary because of random 'noise.' Trial error refers to the nonsystematic error that causes an individual to respond faster or slower to different trials within a given block, independent of prejudice. Trial error might occur because a particular stimulus on a given trial is unusually attention-grabbing or because of momentary distractions (e.g., loud noises, distracting thoughts) in the testing environment.")

[30] See, e.g., Bertram Gawronski et al., Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis, 43 Personality & Soc. Psychol. Bull. 300, 307 (2017) ("we found that individual differences on implicit measures showed significantly lower levels of temporal stability than individual differences on explicit measures"); Jesse Erwin, The IAT: How and When It Works, Psychological Science Observer (Dec. 1, 2007), http://www.psychologicalscience.org/observer/the-iat-how-and-when-it-works#.WJyPIjsrIuU ("The test-retest reliability is okay for research, but not very good for an individual difference measure that you want to be diagnostic of a single person." (quoting Dr. Greenwald, one of the creators of the IAT)).

[31] See, e.g., Russell H. Fazio & Michael A. Olson, Implicit Measures in Social Cognition Research: Their Meaning and Use, 54 Ann. Rev. Psychol. 297, 311 (2003) ("One of the most disturbing trends to emerge in the literature on implicit measures is the many reports of disappointingly low correlations among the measures. . . . In our own lab we have repeatedly failed to observe correlations between IAT measures and priming measures of racial attitudes (r's ranging from −.13 to .05 across four studies)."); Tiffany A. Ito et al., Toward a Comprehensive Understanding of Executive Cognitive Function in Implicit Racial Bias, 108 J. Personality & Soc. Psychol. 187, 208 (2015) ("our analyses revealed that the three measures of racial bias used here showed only weak associations with one another"); Michael A. Olson & Russell H. Fazio, Relations Between Implicit Measures of Prejudice What Are We Measuring?, 14 Psychol. Sci. 636, 638 (2003) ("Confirming our reasoning that they measure different constructs, a traditional version of the BFP and the IAT showed little correspondence.").

[32] See, e.g., Aliza Plener Cover, Hybrid Jury Strikes, 52 HARV. C.R.-C.L. L. REV. 357, 374 (2017) ("As has been demonstrated by social science literature on implicit bias, some jurors may not even be aware of their biases; they may

In fact, trying to identify truly unconscious bias is a daunting task that psychologists have not definitively mastered, and a number of lines of evidence cast doubt on the view that implicit measures actually tap into unconscious bias.[33]

express genuine verbal adherence to egalitarian values while harboring strong yet unconscious biases." (footnote omitted)); Kristen Henning, *Race, Paternalism and the Right to Counsel*, 54 AM. CRIM. L. REV. 649, 651–52 (2017) ("Yet even when individuals stand firmly against racism and are ideologically opposed to explicit bias, they may be impacted by the subconscious influence of implicit bias." (footnote omitted)); Margaret Samadi & Steven D. Rineberg, *Weaving a Stronger Fabric*, 53 TRIAL 48, 50 (2017) ("Implicit bias does not result from any conscious or explicit beliefs, but from unconscious attitudes or stereotypes." (footnote omitted)); Wendell Y. Tong, *Looking Within for Implicit Bias*, 53 TRIAL 27, 28 (2017) ("When we try a case, we need to overcome implicit bias: the unconscious and unspoken emotions and thoughts that affect a juror's or judge's perceptions and decision-making." (footnote omitted)); Leland Ware, *Discriminatory Discipline: The Racial Crisis in America's Public Schools*, 85 UMKC L. REV. 739, 769 (2017); ("It is also important that school personnel take the implicit bias test which will make them aware of their unconscious stereotypes."); *see also supra* text accompanying notes 3–4. *Accord* Bertram Gawronski, Wilhelm Hofmann & Christopher J. Wilbur, *Are ''Implicit'' Attitudes Unconscious?*, 15 CONSCIOUSNESS & COGNITION 485, 486 (2006) ("A widespread assumption underlying the application of indirect measures is that they provide access to unconscious mental associations that are difficult to assess with standard self-report measures." (citations omitted)).

[33] See Blanton & Jaccard, supra note 26, at 278 ("Our review shows that strong inferences regarding the presence and prevalence of unconscious racism are only warranted if one is willing to make strong inferential leaps."); id. at 292 ("Despite researchers' tendency to interpret implicit race data as evidence that unconscious racism is pervasive, logical analysis of this construct and closer inspection of the properties of implicit measures suggest to us that strong conclusions are not warranted at this time."); Gawronski et al., supra note 16, at 496 ("In summary, our review suggests that (a) people may lack awareness of the origin of their attitudes, but that source awareness is not a distinguishing feature of self-reported versus indirectly assessed attitudes, (b) there is no empirical evidence that people lack conscious awareness of indirectly assessed attitudes per se, and (c) there is at least some evidence showing that indirectly assessed (but not self-reported) attitudes can influence other psychological processes outside of conscious awareness. Drawing on these findings, we argue that the term 'unconscious' is adequate for indirectly assessed attitudes only with regard to one particular aspect: impact awareness. However, the term 'unconscious' is inadequate when it is assumed to imply lack of source awareness or lack of content awareness."); Adam Hahn, Charles M. Judd, Holen K. Hirsh & Irene V. Blair, Awareness of Implicit Attitudes, 45 J. Experimental Psych. 1369, 1389 (2014) ("In most academic and popular representations, implicit attitudes are portrayed as "unconscious" and inaccessible to introspection. The current set of studies showed that contrary to this widespread presentation, it is possible to accurately predict the pattern of one's implicit attitudes, without information from a test, even when the implicit attitudes are quite different from explicit

First, recent research shows that many people can accurately predict their pattern of responding on indirect measures of bias.[34] This research suggests that while people may not be able to access in detail the particular structures and processes that give rise to the responses labeled "implicit bias," they are aware of their response tendencies in general (i.e., people may be unaware of the particular sources of a supposed bias but many are aware of the nature of that supposed bias). Unconscious bias as it is commonly discussed within legal domains assumes a lack of awareness of the bias itself, and not just a lack of insight into the precise mechanisms that may give rise to the bias, and this research casts doubt on the view that implicit bias is unknown or unknowable.

Second, people can be effectively instructed on how to fake particular patterns of responding on indirect bias measures such as the IAT, and

feelings toward the same targets, and even when these attitudes might shed a possibly uncomfortable light on a person. In light of these findings, it is important that the characterization of implicit attitudes be carefully considered, both in the academic community and for the general public.").

For expositional purposes only, in this section I treat responses on indirect measures of bias that are interpreted to be evidence of implicit bias as if they are in fact evidence of bias. As suggested in the prior section and as discussed in more detail in the next section, those interpretations are actually quite contestable, often for a large number of persons who have been characterized as possessing an implicit bias based on their scores on the IAT alone. See Hart Blanton, James Jaccard, Erin Strauts, Gregory Mitchell & Philip E. Tetlock, Toward a Meaningful Metric of Implicit Prejudice, 100 J. Applied Psychol. 1468, 1477 (2015) ("The present research found evidence that the zero point of IAT measures designed to assess racial and ethnic prejudices does not map onto empirical estimates of behavioral neutrality. The evidence across the wide range of studies suggests that IAT measures of racial and ethnic prejudice are "right biased" in the sense that, on average, they place scores to the right of a behavioral zero point in a way that implies overestimation of biases in a population.").

[34] See Hahn et al., supra note 33, at 1386 ("We hypothesized that people would be reasonably accurate in their predictions, even when they report very different explicit attitudes. Results from all four studies supported this hypothesis by showing that participants' predictions were considerably accurate under a variety of testing conditions, including one in which participants were given only limited explanation and no experience with the measure before making their predictions (Study 4). We interpret these results to mean that our participants had some awareness of their implicit attitudes—the extent to which they spontaneously respond more positively or negatively toward one target relative to another."); see also Kate A. Ranganath, Colin Tucker Smith & Brian A. Nosek, Distinguishing Automatic and Controlled Components of Attitudes from Direct and Indirect Measurement Methods, 44 J. Experimental Soc. Psychol. 386, 389–92 (2008) (self-reported "gut reactions" and "instant reactions" to gay and straight people correlated with performance on indirect measures of bias); id. at 393 ("These findings suggest that direct measures can be devised to capture automatic components of attitudes despite suggestions that indirect measures are essential for such assessments.").

some people can alter their scores without explicit instruction on how to do so.[35] Thus, people can exert conscious control over the processes that produce the responses to indirect measures of bias in ways that produce patterns of responses in line with their intentions. In addition to raising doubts about whether implicit bias is beyond conscious control, this research cautions against the use of indirect measures of bias as diagnostic tools or screening devices.

Third, presently the empirical data do not allow researchers to declare whether implicit bias and explicit bias are truly distinct.[36] If implicit bias always arose from sources of which people are truly unaware, then there should be no consistent correlation between measures of implicit and explicit bias, but in fact those measures often do correlate.[37]

[35] See, e.g., Klaus Fiedler & Matthias Bluemke, Faking the IAT: Aided and Unaided Response Control on the Implicit Association Tests, 27 Basic & Applied Soc. Psychol. 307, 314–15 (2005) ("The three experiments reported here provide a clear-cut answer to the question of whether controlled responding or faking on the IAT is possible. The answer is an unqualified yes. Moreover, the degree of voluntary manipulation that was possible was quite impressive and was not dependent on explicit instructions of how to accomplish successful faking."); Klaus Fiedler, Claude Messner & Matthis Bluemke, Unresolved Problems with the ''I'', the ''A'', and the ''T'': A Logical and Psychometric Critique of the Implicit Association Test (IAT), 17 Eur. Rev. Soc. Psychol. 74, 98–99 (2006) ("Although respondents are not asked to evaluate attitude targets explicitly, they nevertheless recognise, vividly, that the entire task and the difficulty experienced during sorting is caused by meaningful stimuli, denoting target objects (e.g., Black and White people) and valence cues (pleasant vs unpleasant). Thus, the purpose of testing is not concealed at all. Given this high awareness of the IAT theme, the weaker criterion, lack of voluntary control, would appear to be crucial. However, a recent review suggests that performance on the IAT may be more amenable to voluntary influence than expected. To influence one's IAT score intentionally, it is sufficient to think of admired or disliked individuals or to engage in counter-stereotypical imagery." (citations omitted)); Jessica Röhner et al., What Do Fakers Actually Do to Fake the IAT? An Investigation of Faking Strategies Under Different Faking Conditions, 47 J. Res. Personality 330, 330 (2013) ("Naïve participants successfully faked low scores by slowing down on the congruent block and faked high scores by accelerating on that block.").

[36] Anthony G. Greenwald, T. Andrew Poehlman, Eric Luis Uhlmann & Mahzarin R. Banaji, Understanding and Using the Implicit Association Test: III. Meta-analysis of Predictive Validity, 97 J. Personality & Soc. Psychol. 17, 32 (2009) ("Greenwald and Nosek (2008) have advocated a middle course by treating implicit and explicit measures as empirically distinct constructs, noting that, at present, the question of single versus dual representations appears empirically irresolvable." (citation omitted)).

[37] These correlations are found even across measures that take very different approaches to measurement, and the size of these correlations increases when direct and indirect measures are designed to try to measure the same psychological sources. See B. Keith Payne, Melissa A. Burkley & Mark B. Stokes,

This correspondence suggests that, at least for some people, explicit biases may give rise to implicit biases (i.e., consciously-held beliefs and attitudes may produce the pattern of responses observed on indirect measures of bias) or that the two forms of bias have a common cause.

A related question is whether implicit biases are intentional or unintentional. To the extent intent is synonymous with knowledge or awareness, as it is in many areas of the law,[38] then the evidence of people's insight into their implicit biases could justify labeling implicit biases intentional (at least for those people who can accurately predict their scores on indirect measures of bias). Likewise, to the extent that implicit biases are offshoots of or epiphenomenal to explicit biases, as the evidence of correlations between explicit and implicit bias suggests, one could argue that implicit biases are intentional to the same extent that explicit biases are. But to the extent that intentionality implies a purpose or control over the formation and continuing existence of a bias,[39] then labeling implicit biases intentional is debatable. For many people the associations and/or propositions that that may give rise to implicit biases may be the involuntary product of living in a culture with a history of discrimination and inequality.[40]

Why Do Implicit and Explicit Attitude Tests Diverge? The Role of Structural Fit, 94 J. Personality & Soc. Psychol. 16, 29 (2008) ("We found that tests with poorly matched structures are likely to underestimate the implicit–explicit correlation. But when implicit and explicit tests equate these features, the correlations can be quite high, even on the topic of racial attitudes."); see also Gawronski et al., supra note 16, at 489 ("there is now accumulating evidence that self-reported attitudes are systematically related to indirectly assessed attitudes" (citations omitted)); id. at 490 ("Taken together, the available evidence indicates that self-reported and indirectly assessed attitudes are systematically related. Moreover, the relative size of the correlations seems to depend on a variety of different variables, such as motivational factors, the degree of deliberation during self-report, conceptual correspondence between measures, and measurement error. These findings are in contrast to the widespread assumption that people generally have no conscious access to indirectly assessed attitudes. Rather, it seems that people are consciously aware of the attitudes assessed by indirect measures. However, whether or not these attitudes are reflected in self-report measures depends on a variety of factors pertaining to cognitive, motivational, and methodological variables.").

[38] See David Crump, What Does Intent Mean, 38 Hofstra L. Rev. 1059, 1063–66 (2010) (discussing definitions that equate intent with knowledge or awareness of possible consequences).

[39] See id. at 1062–63 (discussing purpose-based definitions of intent).

[40] See Hal Arkes & Philip E. Tetlock, Attributions of Implicit Prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?", 15 Psychol. Inquiry 257, 274 (2004) ("Assuming that those who participate in the affective priming and IAT methodologies all live in societies in which inequalities exist and are perceived, then as long as the participants are sensitive enough to imbue those inequalities with evaluative significance, nearly everyone will exhibit implicit prejudice and the residues of a racist culture, precisely the results that have been

However, knowledge- and purpose-based definitions of intent typically emphasize awareness of consequences or a purpose to engage in a particular course of conduct. Therefore, the question for intentionality often focuses on an intention to act in a particular way rather than an intention to hold particular views about particular groups. On the one hand, possessing an implicit bias does not mean that one will act on that bias.[41] Therefore, one who holds an implicit bias could justifiably believe that an implicit bias she may possess is unlikely to have adverse consequences. On the other hand, awareness of a bias can prompt efforts to control expression of that bias.[42] Thus, one who engaged in differential treatment of group members with an awareness and ability to control the bias could arguably be said to have engaged in purposeful discrimination. Of course, whether such awareness and control existed would have to be determined on a case-by-case basis, yet that will be very difficult to do outside of experimental settings.

In sum, the extent to which people are consciously aware of and able to control their implicit biases remains an open question. Current evidence indicates, however, that many people are aware of the pattern of responses they are likely to give on indirect measures of bias and can control those response patterns, suggesting a greater degree of consciousness than is often assumed in discussions of implicit bias. To the

reported. Sentient organisms aware of their environment will be accused of harboring hidden biases.").

[41] As discussed in detail in the next section. See infra text accompanying notes 43–58.

[42] See, e.g., Irene V. Blair, The Malleability of Automatic Stereotypes and Prejudice, 6 Personality & Soc. Psychol. Rev. 242, 255 (2002) ("The goal of the present review was to report on the many studies that have directly tested the assumption that automatic stereotypes and prejudice are immutable and inescapable. In contrast to that assumption, the results of these tests show that automatic stereotypes and prejudice can be moderated by a wide variety of events, including, (a) perceivers' motivation to maintain a positive self-image or have positive relationships with others, (b) perceivers' strategic efforts to reduce stereotypes or promote counterstereotypes, (c) perceivers' focus of attention, and (d) contextual cues."); Yi-Wen Chien et al., The Flexible Correction Model: Bias Correction Guided by Naïve Theories of Bias, 8 Soc. & Personality Psychol. Compass 275, 280 (2014) ("A number of studies support the notion that theory-based corrections are guided by both the direction and magnitude of the perceived bias. That is, when asked not to be influenced by a context, people correct in different directions when they hold theories of opposite biases (across targets, across contexts for the same target, or for different people who encounter the same context and target). People correct for perceived biases even when there are no real biases. Corrections for perceived rather than actual bias also mean that people sometimes correct primarily for one perceived bias (i.e. the bias that is most salient or the one associated with a clear theory of bias) even if other biases are operating. People also correct in different amounts when they perceive the bias to be relatively large rather than small." (citations omitted)).

extent people are not aware of their biases, then it may be more difficult to prevent those biases from influencing their behavior. Fortunately, the existing evidence indicates that implicit biases as currently measured and conceptualized do not frequently affect behavior.

IV. IS IMPLICIT BIAS RELATED TO DISCRIMINATORY BEHAVIOR?

Precisely because implicit bias is *implied* or inferred from measures that do not directly assess bias (i.e., indirect bias measures are not, on their face, indisputably measures of bias, whereas direct measures of bias do have facial validity as bias measures if properly designed), one cannot simply look at the responses given on an indirect measure of bias and be sure that the response patterns signify a mental architecture biased against a particular group. A particular response pattern may signify negative associations about one group or sympathetic propositions about that group.[43] The response pattern may be a function of a person's reduced cognitive processing speed due to age, distractions in the testing environment, or some other factor that has nothing to do with group bias per se.[44] The response pattern may be unique to the particular stimuli used on a test, and other stimuli drawn from the same groups may produce different responses because every stimuli potentially activates multiple associations.[45] Further complicating interpretation of responses on indirect measures of bias is the unreliability of these measures: If individuals exhibit different scores on the same or alternative indirect

[43] See supra note 23.

[44] Individual scores on IATs, for instance, are known to be affected by a number of factors that have nothing to do with the associations or propositions one may hold with respect to various groups, such as the order of the tasks used on the test, the age and cognitive ability of the test-taker, and prior practice with a test. See Brian A. Nosek, Anthony G. Greenwald & Mahzarin R. Banaji, The Implicit Association Test at Age 7: A Methodological and Conceptual Review, in Automatic Processes in Social Thinking and Behavior 265, 271–73 (John A. Bargh ed., 2007) (discussing "extraneous influences" on IAT scores).

[45] See, e.g., Blanton & Jaccard, supra note 26, at 280 ("In her study, Devine used priming words that activated unusually negative racial stereotypes, stereotypes that would be rejected by many if they were to encounter and process them consciously. (Her studies used such value-laden words as 'Harlem,' 'prejudice,' 'ghetto,' 'welfare,' 'unemployed,' and 'nigger.'). . . . In fact, research participants respond differently if they are primed with words that are less racist in nature, such as 'black,' "ethnic," 'afro.'"); see also Katie Wolsiefer, Jacob Westfall & Charles M. Judd, Modeling Stimulus Variation in Three Common Implicit Attitude Tasks, 49 Behav. Res. Methods 1193, 1205 ("If the goal of using implicit measures is to assess implicit associations toward attitude objects at a broad, categorical level, it is important to account for the fact that we, as researchers, sample from a population of stimuli just as we sample from a population of students. If we desire to draw inferences past the set of stimuli to which participants respond within an implicit measure, it is important to treat stimuli as random factors in the design.").

measures of bias, which score reflects an individual's true level of implicit bias, or should we even conceive of implicit bias in that way?[46]

Even if we were confident in our interpretation of an individual's responses on an indirect measure of bias, we must still confront the reality that many variables other than just implicit and explicit group biases influence behavior. Many Republicans strongly dislike the Democratic Party and its proposals, but many Republicans have Democrats for friends, and sometimes even vote for Democratic candidates. It is one thing to hold a bias and another to act on that bias.[47]

Because of these interpretation problems and because what the law and organizations mostly care about is whether bias has adverse effects on behavior and outcomes, the key question for implicit bias research from an applied perspective is how implicit bias relates to behavior, if at all. Many studies have examined whether implicit biases (as inferred from indirect measures of bias such as the IAT) predict discriminatory behavior, and the consensus among researchers is (a) that scores on indirect measures of bias (even those scores that are supposedly indicative of a high level of implicit bias) are not reliable predictors of individual behavior and (b) that even in the aggregate (i.e., when we combine the data from many research participants to look for patterns within the aggregate data—an approach that increases the power of the research to find relationships among variables) the correlation between measures of implicit bias and behavior is quite small.

First, with respect to individual-level behavior, the accumulated research findings reveal that it is scientifically inappropriate to use any individual's score on an implicit bias measure as a measure of how likely it is that the individual will have engaged in acts of discrimination in the past or will do so in the future. Indirect measures of bias are presently too unreliable (i.e., they show a high degree of variance across persons, situations, and time) and too poor at predicting individual-level behavior (i.e., they have little ability to predict accurately who will and who will not discriminate in any given situation) to assume that the score reflects

[46] For a discussion of an alternative way of conceptualizing implicit bias—as a dynamic state instead of as a stable thing, as it is often conceived—see Eliot R. Smith & Frederica R. Conrey, Mental Representations are States, not Things: Implications for Implicit and Explicit Measurement, in Implicit Measures of Attitudes 247 (Bernd Wittenbrink & Norbert Schwarz eds., 2007).

[47] This gap between attitudes and behavior exists for both implicit attitudes and explicit attitudes. See Stephen J. Kraus, Attitudes and the Prediction of Behavior: A Meta-analysis of the Empirical Literature, 21 Personality & Soc. Psychol. Bull. 58, 71–72 (1995) ("Today, the attitude behavior relationship is thought of more as a substantive relationship of interest, which will sometimes be large, sometimes be small, and which is influenced significantly by other variables . . . . The question, 'To what extent do attitudes predict future behavior?' is complex and multifaceted and does not readily lend itself to any simple answer . . . .").

the degree to which an individual will or will not discriminate in the future.[48] An individual who supposedly shows high implicit bias on the IAT is no more likely to discriminate in any given situation than an individual who supposedly shows low implicit bias on the IAT. Indeed, in a number of studies "high bias" persons behave *more positively* toward minorities than "low bias" persons.[49] For instance, Lois James and her colleagues conducted a high-fidelity simulation of police-citizen interactions in high-threat situations and found that officers were more cautious in their behavior directed at Black suspects despite being scored as having an implicit bias against African Americans:

> Our police participants demonstrated strong implicit bias associating Black suspects with weapons. This finding is consistent with the psychological literature on racial stereotypes, the experimental research on implicit bias in shooting behavior, and much of the criminological literature on police use of force in the field. However, our participants took *longer* to shoot armed Black suspects than armed White suspects, and they were *less* likely to shoot unarmed Black suspects than unarmed White sus-

[48] See Anthony G. Greenwald, Mahzarin R. Banaji & Brian A. Nosek, Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects, 108 J. Personality & Soc. Psychol. 553, 557 (2015) ("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) and small to moderate predictive validity effect sizes. Therefore, attempts to diagnostically use such measures for individuals risk undesirably high rates of erroneous classifications." (citation and footnote omitted)).

[49] See, e.g., Wendy Berry Mendes & Katrina Koslov, Brittle Smiles: Positive Biases Toward Stigmatized and Outgroup Targets, 142 J. Experimental Psychol. 923, 931 (2013) ("Taken together, these studies show that (a) when individuals have the resources to do so, they display positive biases toward stigmatized and minority group members relative to nonstigmatized or ingroup members; (b) the overcorrection effect appears to be threat based (i.e., the individuals who correct the most tend to be individuals who exhibit higher levels of racial bias or physiological signs of anxiety during social interactions with stigmatized or outgroup members); and (c) by reducing resources, either with stress or cognitive load, the overcorrection effect is attenuated".); J. Nicole Shelton, Jennifer A. Richeson, Jessica Salvatore, & Sophie Trawalter, Ironic Effects of Racial Bias During Interracial Interactions, 16 Psychol. Sci. 397, 400 (2005) ("the higher the White actors' automatic-racial-bias scores, the more positively Black partners perceived them"); María Tortosa, Juan Lupiáñez & María Ruz, Race, Emotion and Trust: An ERP Study, 1494 Brain Res. 44, 48 (2013) ("While race did not affect cooperation rates in the first experiment, the second revealed a bias in favour of black people, and this is so despite the fact that participants showed an IAT effect indicative of a racial bias that favoured whites in both cases.").

> pects. In other words, they were more hesitant and more careful in their decisions to shoot Black suspects. This finding is consistent with our previous experimental research on shooting behavior, some of the criminological literature from police use of force in the field, and qualitative research on officer motivations to shoot or not shoot. Thus, our findings suggest that implicit bias *does not* result in racially motivated decisions to shoot in an expected way—our police participants displayed a counter bias or "reverse racism" effect when tested in a deadly force judgment and decision-making simulator. . . . These findings call into question the validity of the widespread assumption that implicit racial bias is the cause of the disproportionate number of racial minorities in officer-involved shootings.[50]

Because indirect measures of bias cannot be expected to do better than chance at predicting any particular individual's behavior, there is a real risk of mistaken accusations if one equates a person's score on the IAT with a low or high propensity to discriminate. Accordingly, the architects of the IAT have issued a warning against treating individual IAT scores as if they were signs of how likely someone is to discriminate:

> Can (or should) people use this test to make decisions about others? Can one, for example, use this test to measure somebody else's automatic racial preference, and use it to decide that they should or should not serve on a jury? We assert that the IAT should **not** be used in any such way. Especially at this early stage of the IAT's development, it is much preferable to use it mainly to develop awareness of one's own and others' automatic preferences and stereotypes. Using the IAT as the basis

[50] See Lois James, Stephen M. James & Bryan J. Vila, The Reverse Racism Effect: Are Cops More Hesitant to Shoot Black Than White Suspects?, 15 Criminology & Pub. Pol'y 457, 470–71 (2016); see also id. at 469 ("we found that officers were slightly more than three times less likely to shoot unarmed Black suspects than unarmed White suspects"); id. ("Most officers showed moderate (40%) or strong (38%) levels of implicit bias. Perhaps the most relevant finding of the study was that we tested whether IAT scores predicted or were even correlated with decisions to shoot, and we found that they did not, suggesting that implicit bias is unrelated to decisions to shoot in a deadly force judgment and decision-making simulator."). Reviewing the research on whether police officers show a "shooter bias" against minority suspects, Correll and colleagues concluded that "[a]lthough police are affected by target race in some respects, they generally do not show a biased pattern of shooting." Joshua Correll, Sean M. Hudson, Steffanie Guillermo & Debbie S. Ma, The Police Officer's Dilemma: A Decade of Research on Racial Bias in the Decision to Shoot, 8 Soc. & Personality Psychol. Compass 201, 201 (2014).

> for making significant decisions about self or others could lead to undesired and unjustified consequences.[51]

Second, with respect to aggregate-level behavior, meta-analyses of the collected studies conclude that the mean correlation between measures of implicit bias and measures of discrimination is small and variable.[52] In 2013, my colleagues and I synthesized all of the available

[51] *Understanding and Interpreting IAT Results*, PROJECT IMPLICIT, https://implicit.harvard.edu/implicit/demo/background/understanding.html (emphasis in original). Unfortunately, the Project Implicit website, where various IATs can be taken by the general public, continues to give test feedback that is likely to be misleading. In particular, the website characterizes one's results on the IAT as showing strong, moderate, slight, or no bias with respect to a particular group, yet none of those scoring categories has been behaviorally validated: They are based on the researchers' judgment about what constitutes a relative small or large difference in reaction times when responding to stimuli drawn from different groups. And in fact, that judgment has changed over time—the researchers who control Project Implicit website can, by fiat, change their bias labeling conventions and thereby change the supposed distribution of high and low bias among test-takers. These labels have no behavioral meaning: All other things being equal, a person labeled as having a strong automatic preference for Whites over Blacks on the racial attitudes IAT is no more likely to be discriminate in favor of Whites and against Blacks than someone who is labeled as having a strong automatic preference for Blacks over Whites. In short, no behavioral meaning should be attached to the individualized test feedback given on Project Implicit. For a fuller discussion of the problems with the scoring categories used with the IAT, see Blanton & Jaccard, *supra* note 26, at 290–92; Blanton et al., *supra* note 33, at 1469.

[52] Most social science studies, including implicit bias studies, aggregate data over research participants and then examine the degree of relationship between two or more variables within the aggregate data set. For instance, studies into the relation between implicit bias and behavior will administer an indirect measure of bias toward particular groups (most often Whites and Blacks) and then examine how these participants behave toward members of those groups. Data reflecting individual bias scores and individual's behaviors will then be aggregated and the degree of overall relationship between bias and behavior within that aggregate dataset will be assessed. These studies yield correlation coefficients that reflect the strength of the relationship between bias and behavior. A perfect relationship (e.g., if every person receiving a no-bias score showed no discrimination and every person receiving a bias score showed discrimination) would produce a correlation coefficient of 1, and no relationship (e.g., half the persons with no-bias scores discriminate and half with the bias scores discriminate) would produce a correlation coefficient of 0. Assuming that the study's findings would generalize to other samples, then the study would permit an inference about the relationship of bias and behavior at the aggregate level only (correlations reflect patterns in aggregate data and are not measures of the relation between bias and behavior for particular individuals). See Eric Luis Uhlmann et al., Getting Explicit About the Implicit: A Taxonomy of Implicit Measures and Guide for Their Use in Organizational Research, 15

IAT studies examining the correlation between measures of implicit racial and ethnic bias and measures of discriminatory behavior, and we found that the average correlation between measures of implicit racial bias and behavior was only .15 and between measures of implicit ethnic bias and behavior was only .12.[53] As notable as the small size of the average correlations was the accompanying large variance observed in

Organizational Res. Methods 553, 582 (2012) ("It should be noted that even when they correlate with behavior, implicit measures typically rely on arbitrary metrics inappropriate for individual diagnostic assessment. For instance, that a person scores a þ.5 on a race IAT (a score indicating a stronger pattern of association between White and good and Black and bad) is not independently informative about the individual. Rather, this value is only meaningful in the context of a greater data set, and only for prediction." (citation omitted)).

[53] *See* Frederick Oswald, Gregory Mitchell, Hart Blanton, James Jaccard & Philip E. Tetlock, *Predicting Ethnic and Racial Discrimination: A Meta-analysis of IAT Criterion Studies*, 105 J. Personality & Soc. Psychol. 171, 182 (2013) ("Our meta-analytic estimates of the mean correlations between IAT scores and criterion measures of racial and ethnic discrimination are smaller than analogous correlations reported by Greenwald, Poehlman, et al. (2009): overall correlations of .15 and .12 for racial and interethnic behavior compared to correlations of .24 and .20 for racial and other intergroup behavior reported by Greenwald and colleagues." (citation omitted)). The creators of the IAT have agreed that the correlations between IAT scores and behavioral measures are small, though they argued for a slightly higher estimate than we found. *See* Greenwald et al., *supra* note 48, at 560 ("both studies agreed that, when considering only findings for which there is theoretical reason to expect positive correlations, the predictive validity of Black–White race IATs is approximately r = .20"). They speculate that these small effects may accumulate to produce noticeable disparities in treatment of different groups, but as we discussed in response, their speculations raise numerous debatable points and presently lack empirical backing. *See* Frederick Oswald, Gregory Mitchell, Hart Blanton, James Jaccard & Philip E. Tetlock, *Predicting Ethnic and Racial Discrimination with the IAT: Small Effect Sizes of Unknown Societal Significance*, 108 J. Personality & Soc. Psychol. 562, 569 (2015) ("Assertions about the cumulative effects of small effect sizes should be counted not as evidence but as starting points for future efforts to identify substantively and theoretically important moderator variables and boundary conditions."); *see also id.* at 562 ("No matter which data selection rules were followed, no matter how the data were aggregated, and no matter which statistical approach was employed to analyze the data, mean effect sizes within and across data groupings generally were small (or very small) and often not in line with theoretical predictions or commonsense expectations. Nothing presented in the reanalysis of our meta-analysis by Greenwald, Banaji, and Nosek (2015) alters that conclusion. This convergence of findings by two different research groups indicates that, by current scientific standards, IATs possess only a limited ability to predict ethnic and racial discrimination and, by implication, to explain discrimination by attributing it to unconscious biases." (citation omitted)).

these correlations across studies.[54] In a number of studies, the observed correlation was negative (i.e., high bias was related to less discriminatory behavior) or approximately zero (no correlation at all was observed). For instance, when we examined the degree to which the IAT could predict behaviors indicative of in-group favoritism (e.g., pro-White behavior by someone who supposedly holds an implicit bias *in favor of* Whites), we found no correlation between in-group implicit bias and acts of in-group favoritism.[55] Likewise, we found that measures of implicit bias were no better than measures of explicit bias at predicting spontaneous behaviors (e.g., performance on a video game task meant to simulate police decisions to shoot or not shoot an assailant) or "microaggressions" (e.g., amount of smiling at an interacting partner or how close one sits to an interacting partner).[56]

A recent meta-analysis conducted by Brian Nosek (one of the IAT architects) and his colleagues examined even more studies than we did and concluded that the average correlation between measures of implicit bias and behavior was smaller than the correlation we observed (r = .10).[57] A correlation this size means that if we had implicit bias scores for 100 persons, and those persons made 100 decisions, only one or two of those decisions might be different if we were to draw a different distribution and compare the decisions of the two groups (i.e., only approximately 1% of the variance in observed behavior can be explained by scores on an indirect measure of bias). This small correlation means that only by examining the behavior of a large group of persons might we find even a few decisions within this group that were possibly influenced by implicit bias.

In sum, the empirical evidence now establishes that indirect measures of bias such as the IAT should not be interpreted as measures of an individual's propensity to discriminate. Even when we aggregate data from many individuals, the correlation between indirect measures of bias and measures of discriminatory behavior is low and highly variable, often producing results that are inconsistent with interpretations of indirect measures of bias as valid measures of bias.

[54] See Oswald et al. (2013), supra note 53, at 185 ("The tremendous heterogeneity observed within and across criterion categories indicates that how implicit biases translate into behavior—if and when they do at all—appears to be complex and hard to predict.").

[55] Id. at 186 ("[F]or the interpersonal behavior and person perception criteria, the race IAT was a poor predictor of behavior toward Whites. [Implicit-criterion correlations] were close to zero or negative for White-target-only criteria other than brain activity.").

[56] Id. at 184 ("[T]he race and ethnicity IATs were weak or unreliable predictors of the more spontaneous behaviors covered by this meta-analysis.").

[57] See Patrick S. Forscher et al., A Meta-analysis of Change in Implicit Bias 36 (under submission), https://osf.io/b5m97/.

These behavioral data bear special emphasis in light of the common claim in popular discussions of implicit bias that most people hold implicit racial and gender biases that may lead to adverse behaviors.[58] The behavioral data make clear that people who are scored as being implicitly biased are *not* more likely to engage in an act of discrimination than those whose scores on indirect measures are interpreted as evidence of no implicit bias. Moreover, if we have two corporations identical in all respects except that one corporation has a higher percentage of persons with implicit bias than the other, we *cannot* validly predict that there will many, if any, differences in the personnel decisions made within these two corporations and we *cannot* identify any particular persons more or less likely to discriminate from any measure of implicit bias. Implicit bias, as it is presently conceived and measured, is not a good predictor of behavior by individuals or groups.

V. CAN IMPLICIT BIAS BE CHANGED THROUGH TRAINING OR EDUCATION?

In light of uncertainties about what exactly indirect measures of bias measure and in light of the evidence that, whatever that is, it is only weakly related to behavior, one might ask whether it is wise to devote resources to changing implicit bias until we have a better understanding of its sources and effects. Nonetheless, many organizations and institutions are now devoting considerable resources to the problem of implicit bias and commissioning consultants to develop and administer programs aimed at raising awareness about implicit bias, reducing implicit bias, and preventing implicit bias from affecting behavior.[59]

[58] See, e.g., Banaji & Greenwald, supra note 5, at 189–209 (Appendix 2, linking racial disparities to racial discrimination). Very few implicit bias studies have actually employed nationally representative samples that would allow such estimates. Instead, data collected through voluntary participation on the Project Implicit website serves as the primary basis for these claims.

[59] See, e.g., Ellen Huett, Rise of the Bias Busters: How Unconscious Bias Became Silicon Valley's Newest Target, Forbes (Nov. 2, 2015, 10:00 AM), https://www.forbes.com/sites/ellenhuet/2015/11/02/rise-of-the-bias-busters-how-unconscious-bias-became-silicon-valleys-newest-target/#1c482f3d19b5 ("The idea that you can methodically stamp out unconscious bias (also called implicit or hidden bias) has caught fire with tech companies because it's relatively new, data-driven and blameless – everyone is told they have it and can't avoid it, so no one is singled out or gets defensive."); Joann S. Lublin, Bringing Hidden Biases Into the Light, Wall Street J. (Jan. 9, 2014), https://www.wsj.com/articles/bringing-hidden-biases-into-the-light-1389311814 ("As they struggle to diversify their workforces, big businesses are teaching staffers to recognize that 'unconscious bias'—or an implicit preference for certain groups—often influences important workplace decisions."). Many of the programs being marketed as implicit bias interventions have never been subjected to rigorous testing for their effects on bias or behavior. See, e.g., Gregory Mitchell, Jumping to Conclusions: Advocacy and Application of Psychological

The likelihood that these programs will have any direct effect on bias or behavior is small because even programs designed by psychologists to target the mechanisms that may underlie implicit bias have not been very effective. Programs that attempt to change the associations or propositions that one holds with respect to members of a particular group and that make anti-discrimination norms salient right before administration of an indirect measure of bias can affect the score subsequently observed on that measure.[60] But very few studies have examined the long-term effects of such interventions by monitoring changes in implicit bias measurements at multiple points in time following the interventions.[61] Most notably, the interventions have been found to have little or no effect on behavior (i.e., have not been shown to change how people behave toward members of other groups).[62] This fact may reflect the tenuous relationship between implicit bias and behavior just as much as the ineffectiveness of the interventions.

Notwithstanding the lack of evidence that implicit bias interventions directly prevent discrimination, to the extent such programs increase vigilance about behavior and encourage scrutiny of outcomes for possible group-based disparities, then the programs may indirectly provide a benefit to an organization and its members.[63] However, that possible indi-

Research, in The Politics of Social Psychology 139, 147–48 (Lee Jussim & Jarret T. Crawford eds., 2017) (discussing implicit bias program targeted at police departments).

[60] See Forscher et al., supra note 57, at 33 ("We found that implicit bias can be changed across many areas of study, populations, implicit measures, and research designs. The type of approach used to change implicit bias mattered greatly. Some procedures were effective at changing implicit bias, whereas others were not."); id. at 34 ("However, even the procedures that produced robust effects on implicit bias had 'small' effect sizes by conventional standards and as compared to typical effect sizes in social psychology. Our funnel plot analyses suggest that the true population effects of these procedures may be even smaller than our meta-analytic estimates due to publication bias, p-hacking, and/or other related processes." (citations omitted)).

[61] See id. ("Another limit to generalizability is a lack of research interest in change beyond the confines of a single experimental session. Only 17 (3.0%) samples used procedures that took longer than one session to complete. Only 38 (6.6%) samples in the meta-analysis collected longitudinal outcomes and therefore had the opportunity to examine whether the procedures they investigated produce long-term changes. Short-term changes in implicit bias do not necessarily generalize to longer-term changes." (citations omitted)).

[62] What changes in behavior were observed did not appear to bear any relation to changes in implicit bias: "We found [] little evidence that procedures that change implicit bias also produce change in behavior. We also found no evidence that changes in implicit bias mediate changes in behavior, nor that changes in behavior mediate changes in implicit bias." Id. at 37.

[63] Cf. Patrick S. Forscher et al., Breaking the Prejudice Habit: Mechanisms, Timecourse, and Longevity, 72 J. Experimental Soc. Psychol. 133, 144 (2017)

rect benefit should be weighed against the potential opportunity costs if adoption of the implicit bias program crowds out other interventions that may more directly or more effectively prevent discrimination. Furthermore, adoption of an implicit bias intervention should be considered in light of the possible adverse effects that may come with such a program. These potential costs include: (a) loss of trust among employees and management due to increased suspicion that bias contaminates decisions and due to stigmatizing managers and employees as persons likely to engage in discrimination,[64] (b) more impersonal and uncomfortable interactions among members of different groups,[65] and (c) compensatory

(arguing that an intervention had important positive effects despite producing lasting change in implicit bias: "Overall, our results suggest that the habit-breaking intervention does not primarily exert its effects through the strategies the participants learn, nor does it exert its effects by changing the raw quantity of race-related thoughts, race-related conversations, or interracial interactions. Instead, the habit-breaking intervention increases people's sensitivity to bias, particularly when others act with bias, and increases the probability that, when a person encounters bias, he or she will label that bias as wrong. The process of detecting bias in others and labeling it as wrong may in turn provoke concern about racial discrimination, just as the concern may itself provoke the detection of future bias.").

[64] See Katherine T. Bartlett, Making Good on Good Intentions: The Critical Role of Motivation in Reducing Implicit Workplace Discrimination, 95 Va. L. Rev. 1893, 1959 (2009) ("Imposing surveillance regimes on all workplaces ignores the impact of excessive suspicion and micromanagement on the kind of trust, organizational esprit, and commitment to nondiscrimination norms." (footnote omitted)); William T. Self, Gregory Mitchell, Barbara A. Mellers, Philip E. Tetlock, & J.A.D. Hildreth, Balancing Fairness and Efficiency: The Impact of Identity-Blind and Identity-Conscious Accountability on Applicant Screening 13, Plos One (2015) ("[P]articipants reacted negatively to [identity-conscious] accountability instructions, reporting feelings of less trust in their decision-making ability. These results suggest that strong forms of outcome accountability may impose a social cost on the organization and have negative implications for the psychological contract between managers and the organization."),
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0145208.

[65] Many implicit bias training sessions encourage participants to complete an IAT as part of the training, yet taking an Implicit Association Test has been found to produce worse intergroup interactions because people who fear they may be implicitly biased can become more cautious and inhibited during their interactions with persons from other groups. See Jacquie D. Vorauer, Completing the Implicit Association Test Reduces Positive Intergroup Interaction Behavior, 23 Psychol. Sci. 1168, 1173 (2012) ("The results of these studies clearly indicate that being alerted to potential bias and limited response control through taking the IAT has negative implications for White individuals' intergroup interaction behavior."). More generally, anxiety about appearing prejudiced can adversely affect the quality or quantity of intergroup interactions. See, e.g., E. Ashby Plant & Patricia G. Devine, The Antecedents and Implications of Interra-

measures aimed at promoting particular groups at the expense of merit-based decision-making.[66]

CONCLUSION: WHERE FROM HERE?

On a number of important points, there is now consensus among psychologists that contradicts common beliefs about implicit bias. Individual scores on measures of implicit bias are unreliable and do not reflect an individual's tendency to discriminate. Implicit bias as it is presently being assessed bears only a weak relationship to behavior in the collected studies. And current implicit bias interventions have no proven long-term effects on bias or behavior.

On a number of other points, there is now uncertainty among psychologists that undercuts common beliefs about implicit bias. Contrary to common portrayals, indirect measures of bias such as the IAT do not necessarily tap into mental associations people have with respect to various groups. How mental associations and propositions may interact to influence the expression of bias is unclear. Implicit bias and explicit bias may or may not have a common source. And implicit bias is not always an unconscious bias, though it may be for some people.

Many of the confusions, errors, and omissions encountered in popular portrayals of implicit bias research can probably be explained by the complicated nature of the topic and the changing empirical terrain. First, the primary research literature can be difficult to access and difficult to understand, for it often appears in specialized journals, is laden with jargon and technical details, and often presumes prior knowledge of research topics. Second, many of the popular accounts of implicit bias research have relied on articles written by psychologists that were aimed at translating this technical literature for the general public, yet many of

cial Anxiety, 29 Personality & Soc. Psychol. Bull. 790, 798 (2003) ("The findings across both studies also supported the argument that anxiety regarding interracial interactions is associated with avoidance of such interactions and, in Study 1, hostility regarding interracial interactions. In addition, in Study 2, anxiety regarding an upcoming interracial interaction resulted in people being less likely to participate in the interaction. This finding is important because it establishes that, consistent with the model and previous theorizing, intergroup anxiety actually creates the tendency to avoid interracial interactions. The fact that those who experience anxiety regarding interracial interactions are more likely to avoid such interactions is particularly problematic for the likelihood of improving intergroup relations." (citations omitted)).

[66] See Self et al., supra note 64, at 13 ("[T]he pro-female and pro-black biases observed under [identity-conscious] accountability conditions resulted in the selection of less-qualified candidates across labor pools. These results are consistent with research suggesting that accountability instructions that focus decision-makers on outcomes, as opposed to processes, result in more cursory information processing and a shift in attention from job-relevant to job-irrelevant factors.").

these articles were written before important developments occurred within this body of research. Indeed, these first-generation articles often extrapolated from limited data to make strong claims or predictions that were not borne out over time.[67] Thus, some popular accounts may reflect a fair understanding of the state of play in the implicit bias research field as it then existed.

The concern that unconscious biases may contaminate many decision-making processes and cause spontaneous negative reactions in intergroup interactions, and yet escape detection precisely because those biases are unconscious, motivates many researchers to try to find ways to identify unconscious biases and prevent their adverse effects. Present attempts to measure unconscious bias have proved less successful than originally hoped and as often portrayed in the law reviews and popular press. The bias inferred from indirect measures of bias—that is, implicit bias—may or may not reflect unconscious biases, but whatever is being measured does not reliably predict behavior, whether that behavior is subtle, spontaneous, or deliberate.

The challenge for implicit bias researchers now, and for those chronicling that research to others, is to update our understanding of what implicit bias may be and how it may relate to behavior in the real world. This updated understanding should produce greater reluctance among social observers and commentators to attribute harmful acts against women and minorities to unconscious bias. The all-too-common incidents of police shootings of minorities and disrespect toward women often prompt invocations of unconscious bias as an explanation, yet explicit (i.e., consciously accessible) bias is just as likely a cause if group membership was in fact a motivating factor.[68] Because unconscious bias

[67] For example, early on implicit bias researchers repeatedly claimed that implicit bias measures would turn out to be much better predictors of behavior than explicit bias measures because the former avoided social pressure concerns and did not require introspective access to biased thoughts. Yet that prediction has turned out to be false. For a fuller recounting of the history of implicit bias research and its public dissemination, see Gregory Mitchell & Philip E. Tetlock, *Popularity as a Poor Proxy for Utility: The Case of Implicit Prejudice*, *in* PSYCHOLOGICAL SCIENCE UNDER SCRUTINY: RECENT CHALLENGES AND PROPOSED SOLUTIONS 164, 166–72 (Scott O. Lilienfeld & Irwin D. Waldman eds., 2017).

[68] The correlations observed between explicit bias measures and measures of discriminatory behavior tend to be the same size as those observed with respect to implicit bias measures and behavior, or even slightly larger for some biases. See Oswald et al. (2013), supra note 53, at 183 ("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."). Of

by its nature is not something whose presence and effects can be directly observed at the time an incident occurs, post hoc attempts to assign causation to unconscious bias are inherently flawed from a scientific perspective. Flawed applications of implicit bias research to particular cases cannot be a sound basis for assigning blame. Likewise, so long as we accept vague attributions of societal inequalities to implicit bias, rather than specify and test particular causal pathways by which implicit bias may have produced observed differences in education, employment, policing, and health care, we will continue to spend resources on unproven programs and squander opportunities for effective change.

We can learn from the limitations and difficulties that have been revealed by the second generation of implicit bias research and perhaps eventually develop a better understanding of the sources of discrimination and new ways to reduce its occurrence. But flawed understandings of what implicit bias research has found cannot be a sound basis for public policy.

course, in many situations, it will be difficult to determine whether group affiliations really did play a causal role.

***