California Law Review

VOL. 94 JULY 2006 NO. 4

Implicit Bias: Scientific Foundations

Anthony G. Greenwald†
Linda Hamilton Krieger††

The assumption that human behavior is largely under conscious control has taken a theoretical battering in recent years. Although this assault in some ways resembles the previous century's Freudian revolution, there are important differences between the two. Freud's views of unconscious mechanisms were embedded in a theory that never achieved conclusive support among scientists, despite many empirical theory-testing efforts in the middle third of the twentieth century.[1] Consequently, most psychologists have abandoned Freud's psychoanalytic theory of unconscious mental processes.

Theoretical conceptions of conscious control over human behavior were strongly re-established in the last third of the twentieth century, but the dominance of such views has been crumbling during the past two decades. Unlike the Freudian revolution, however, the new science of

† Professor of Psychology, University of Washington.

†† Professor of Law, University of California, Berkeley School of Law (Boalt Hall). Thanks to Jennifer Eberhardt, Jerry Kang, Tom Newkirk, and Jeff Rachlinski for comments on preliminary versions of this article, and to Ian Ayres, Gary Blasi, Jack Dovidio, John Jost, and Mahzarin Banaji for useful discussions that preceded the writing.

1. *See* MATTHEW HUGH ERDELYI, PSYCHOANALYSIS: FREUD'S COGNITIVE PSYCHOLOGY (1985); Matthew Hugh Erdelyi & Benjamin Goldberg, *Let's Not Sweep Repression Under the Rug: Toward a Cognitive Psychology of Repression*, *in* FUNCTIONAL DISORDERS OF MEMORY 355 (John F. Kihlstrom & Frederick J. Evans eds., 1979); Anthony G. Greenwald, *New Look 3: Unconscious Cognition Reclaimed*, 47 AM. PSYCHOL. 766 (1992); John F. Kihlstrom, *The Psychological Unconscious*, *in* HANDBOOK OF PERSONALITY: THEORY AND RESEARCH 445 (Lawrence A. Pervin ed., 1990); Howard Shevrin & Scott Dickman, *The Psychological Unconscious: A Necessary Assumption for All Psychological Theory?*, 35 AM. PSYCHOL. 421 (1980).

unconscious mental process is not the product of a single brilliant theoretical mind. Rather, it is being constructed from an evolving, accumulating body of reproducible research findings.[2]

This Article introduces *implicit bias*—an aspect of the new science of unconscious mental processes that has substantial bearing on discrimination law. Theories of implicit bias contrast with the "naïve" psychological conception of social behavior,[3] which views human actors as being guided solely by their explicit beliefs and their conscious intentions to act. A belief is *explicit* if it is consciously endorsed. An intention to act is *conscious* if the actor is aware of taking an action for a particular reason. Of course, actors may dissemble and deny they are taking an action for a particular reason, so conscious intentions based on explicit beliefs may be hard to verify. But a deceitful actor is nevertheless *capable* of asserting the belief or identifying the intention that provides the basis for action, even when *unwilling* to do so.[4] In contrast, the science of implicit cognition suggests that actors do not always have conscious, intentional control over the processes of social perception, impression formation, and judgment that motivate their actions.

---

2. The early stages of this modern revolution are reviewed by Greenwald, *supra* note 1. Nisbett and Wilson's exposé of the inadequacies of introspective explanations of behavior was a noticeable starting point of the modern revolution, leading to widespread understanding that the self-report measures of conscious mental process that were widely used in psychological research were highly suspect. *See* Richard. E. Nisbett & Timothy DeCamp Wilson, *Telling More Than We Can Know: Verbal Reports on Mental Processes*, 84 PSYCHOL. REV. 231 (1977). Wegner's and Bargh's more recent works reveal the frequency with which seemingly ordinary voluntary actions are controlled in ways that evade conscious scrutiny, further undermining the idea that a conscious mind is the effective governor of most human behavior. *See generally* DANIEL M. WEGNER, THE ILLUSION OF CONSCIOUS WILL (2002); John A. Bargh et al., *The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals*, 81 J. PERSONALITY & SOC. PSYCHOL. 1014 (2001).

3. "Naive psychology" refers to laypersons' intuitions about determinants and consequences of human thought and behavior, especially their own. Modern treatments were largely inspired by Fritz Heider's book, *The Psychology of Interpersonal Relations*, which initiated systematic investigation of how laypersons' intuitions differ from scientific understanding. FRITZ HEIDER, THE PSYCHOLOGY OF INTERPERSONAL RELATIONS (1958).

4. Methodological investigations by social psychologists in the 1960s revealed social influences operating within research and interview settings that would lead people to describe their explicit beliefs inaccurately in experimental studies. *See* Martin T. Orne, *On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications*, 17 AM. PSYCHOL. 776 (1962); Milton J. Rosenberg, *The Conditions and Consequences of Evaluation Apprehension*, *in* ARTIFACT IN BEHAVIORAL RESEARCH 279 (Robert Rosenthal & Ralph L. Rosnow eds., 1969); Stephen J. Weber & Thomas D. Cook, *Subject Effects in Laboratory Research: An Examination of Subject Roles, Demand Characteristics, and Valid Inference*, 77 PSYCHOL. BULL. 273 (1972). Work inspired by Festinger's cognitive dissonance theory initiated modern interest in understanding people's inability to identify the causes of their own thought and behavior. *See* LEON FESTINGER, A THEORY OF COGNITIVE DISSONANCE (1957). Nisbett and Wilson's article summarizes the humbling implications of the ensuing two decades of research. *See* Nisbett & Wilson, *supra* note 2.

I
IMPLICIT COGNITION

Many mental processes function implicitly, or outside conscious attentional focus.[5] These processes include implicit memory,[6] implicit perception,[7] implicit attitudes,[8] implicit stereotypes,[9] implicit self-esteem,[10] and implicit self-concept.[11] The meaning of *implicit* in these phrases is technical, but still reasonably close to its everyday meaning. For example, research on "implicit memory" demonstrates that even when a person cannot voluntarily ("explicitly") retrieve a memory, that person's behavior may reveal that some previous experience has left a memory record. In such situations, the memory is said to be expressed implicitly, and not explicitly, in the behavior. For example, on the first day of one implicit-memory experiment,[12] subjects were asked to pronounce each of a long list of people's names. Some of these names were recognizably famous, while others were not. On Day Two, these same subjects judged whether each name on another long list was famous or not. Half of Day Two's non-famous names were repeated from Day One. The result: On Day Two, more of the *repeated* non-famous names than the novel ones were judged famous. These "false fame" judgments comprise an implicit-memory effect. The names acquired some familiarity from Day One's attended-but-not-studied pronunciation even though, by Day Two, the subject often did not consciously remember the initial exposure on Day One. This perhaps vague feeling of familiarity for repeated names was sometimes misattributed to fame, leading to greater false judgments of fame for the repeated than the non-repeated names. Subjects presumably go through a mental

5. For an overview of implicit social cognition, which encompasses the phenomena of implicit attitudes, stereotypes, self-esteem, and self-concept, see Anthony G. Greenwald & Mahzarin R. Banaji, *Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes*, 102 PSYCHOL. REV. 4 (1995); *see also* Anthony G. Greenwald et al., *A Unified Theory of Implicit Attitudes, Stereotypes, Self-Esteem, and Self-Concept*, 109 PSYCHOL. REV 3 (2002).

6. *See* Daniel L. Schacter, *Implicit Memory: History and Current Status*, 13 J. EXPERIMENTAL PSYCHOL.: LEARNING, MEMORY, & COGNITION 501 (1987); Larry L. Jacoby & Mark Dallas, *On the Relationship Between Autobiographical Memory and Perceptual Learning*, 110 J. EXPERIMENTAL PSYCHOL: GENERAL 306 (1981).

7. *See* John F. Kihlstrom et al., *Implicit Perception*, *in* PERCEPTION WITHOUT AWARENESS: COGNITIVE, CLINICAL, AND SOCIAL PERSPECTIVES 17 (Robert F. Bornstein & Thane S. Pittman eds., 1992).

8. *See* Anthony G. Greenwald et al., *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, 74 J. PERSONALITY & SOC. PSYCHOL 1464 (1998).

9. *See* Laurie A. Rudman et al., *Implicit Self-Concept and Evaluative Implicit Gender Stereotypes: Self and Ingroup Share Desirable Traits*, 27 PERSONALITY & SOC. PSYCHOL. BULL. 1164 (2001).

10. *See* Anthony G. Greenwald & Shelly D. Farnham, *Using the Implicit Association Test to Measure Self-Esteem and Self-Concept*, 79 J. PERSONALITY & SOC. PSYCHOL 1022 (2000).

11. *See* Rudman et al., *supra* note 9.

12. *See* Larry L. Jacoby et al., *Becoming Famous Overnight: Limits on the Ability to Avoid Unconscious Influences of the Past*, 56 J. PERSONALITY & SOC. PSYCHOL. 326 (1989).

process resembling the following: "This name seems familiar. Why is it familiar? Perhaps it's famous." For names that subjects explicitly remembered seeing and pronouncing on Day One, subjects correctly understood why the name seemed familiar. Therefore they did not mistakenly attribute the familiarity to fame.

II
IMPLICIT ATTITUDES AND IMPLICIT STEREOTYPES

Implicit-memory research conducted in the 1980s led researchers to develop measures of other implicit mental phenomena. Two of these—*implicit attitudes* and *implicit stereotypes*—are especially relevant to bias and discrimination.

A. Implicit Attitudes

Social psychologists define an *attitude* as an evaluative disposition—that is, the tendency to like or dislike, or to act favorably or unfavorably toward, someone or something. Explicit expressions of attitudes occur frequently, whenever we say we like or dislike someone or something. A statement that one likes a particular presidential candidate provides a ready example. Attitudes can also be expressed through favorable or unfavorable *action*, such as by voting for or against a particular presidential candidate. If the voter understands that the favorable vote results from favorable beliefs about the candidate, the vote is an *explicit attitude expression*.

In other situations, a vote might function as an *implicit attitude indicator*—that is, an action that indicates favor or disfavor toward some object but is not understood by the actor as expressing that attitude.[13] For example, a voter may vote for a particular candidate even though the voter knows nothing other than the candidate's name. One of the things that might influence a voter to vote for this candidate is that the candidate's name shares one or more initial letters with the voter's name. In such a case, the vote can be understood, at least in part, as an implicit expression of the voter's self-favorable attitude.[14]

As an additional, hypothetical example, consider how people form impressions of a liked or disliked candidate's spouse, child, or sibling. Someone who knows nothing about the candidate's relative other than the relative's relation to the candidate may find that they like or dislike the relative. Not surprisingly, this attitude toward the relative is likely to match the attitude toward the candidate. Evaluation of the unknown relative may therefore be regarded as an *implicit indicator* of attitude toward the

13. Greenwald and Banaji define implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects." Greenwald & Banaji, *supra* note 5, at 8.

14. *See id.* at 12.

candidate. Here, the "implicit" designation indicates that the attitude expressed toward the candidate determined the attitude toward the relative, even though the liking or disliking for the relative may be experienced as an independent attitude.

Implicit attitudes are of greatest interest when implicit and explicit attitudes toward the same object differ. These discrepancies are referred to as *dissociations* between implicit and explicit attitudes. Dissociations are commonly observed in attitudes toward stigmatized groups, including groups defined by race, age, ethnicity, disability, and sexual orientation. Researchers have used the Implicit Association Test (IAT), a procedure described below, to reveal such attitudinal dissociations.[15]

B. Implicit Stereotypes

A social *stereotype* is a mental association between a social group or category and a trait. The association may reflect a statistical reality, but it need not. If the association does reflect a statistical reality, members of the group will be more likely to display the trait than will members of other groups. A perfect or near-perfect correlation, which might be a *defining trait*—such as physical stamina for basketball players—is of little psychological interest and is often not even considered part of a stereotype. It is of greater psychological interest when the correlation between group membership and trait expression is much less than perfect, but the trait nevertheless distinguishes members of one group from others. Suppose, hypothetically, that 10-15% of people over the age of seventy drive on highways at speeds noticeably below speed limits, but that only 5% of younger people drive this slowly. If these figures were accurate, they would reflect a genuine association between age and driving behavior. However, the stereotype would apply to only a small minority (10-15%) of elderly people. Nonetheless, it may come to serve as a default assumption—the assumption that *any* elderly person is likely to drive slowly.

The first experimental demonstration of implicit stereotypes made use of the stereotype that associates male gender with fame-deserving achievement.[16] In this experiment, which was based on Jacoby et al.'s false-fame implicit-memory effect described in Part I,[17] Banaji and Greenwald found that the false-fame effect was substantial when the pronounced (i.e., attended-but-not-studied) names were male, but was

---

15. *See* Wilhelm Hofmann et al., *A Meta-Analysis on the Correlation Between the Implicit Association Test and Explicit Self-Report Measures*, 31 PERSONALITY & SOC. PSYCHOL. BULL. 1369 (2005) (discussing factors that promote dissociation of implicit from explicit attitudes); Brian A. Nosek, *Moderators of the Relationship Between Implicit and Explicit Evaluation*, 134 J. EXPERIMENTAL PSYCHOL.: GENERAL 565 (2005).

16. *See* Mahzarin R. Banaji & Anthony G. Greenwald, *Implicit Gender Stereotyping in Judgments of Fame*, 68 J. PERSONALITY & SOC. PSYCHOL. 181 (1995).

17. *See* Jacoby et al., *supra* note 12.

noticeably weaker when the names were female. Banaji and Greenwald described this result as an implicit indicator of the stereotype that associates maleness with fame-deserving achievement.[18] Put more technically, an implicit stereotype of this kind can be defined as "the introspectively unidentified (or inaccurately identified) traces of past experience that mediate attributions of qualities to members of a social category."[19]

Stereotypes can involve associations of either favorable or unfavorable traits with a group. Because the favorable-unfavorable distinction is also central to the concept of attitude, it is natural to ask how stereotypes and attitudes differ. For stereotypes, the content of the ascribed trait, rather than its evaluative valence, is central. For attitudes, the opposite holds. For example, in the implicit-fame experiment, it was the *trait* of fame, rather than the *positivity* of fame, that defined the implicit-stereotype phenomenon.

III
RESPONSE BIAS AND IMPLICIT BIAS

The term "bias," sometimes referred to as "response bias," denotes a displacement of people's responses along a continuum of possible judgments. Response bias need not indicate something unwise, inappropriate, or even inaccurate. For example, instructors may vary in their response bias in grading, such that some assign a relatively high grade to average student performance while others assign a lower grade to the same performance. Instructors who differ in response bias on the grading dimension may nevertheless be equally sensitive to differences among students. Consider an instructor who is biased to grade leniently and assigns grades exclusively between A (highest) and C (lowest). This instructor's grades may be perfectly correlated with those of a severe-grading instructor who limits grades to the B-to-D range. If these two instructors graded the same work, each of the lenient instructor's grades would be exactly one letter grade above those of the more severe instructor. Unless there are established standards that associate specific performances with specific grades, one could not accuse either instructor of being less "accurate" than the other.

A more widely recognized form of bias does affect response accuracy and bears a pejorative connotation. Imagine a particular instructor who differentially assigns grades to two identically performing students when one student is male and the other is female, or when one is White and the other is Black. In this case, the fairness and accuracy of judgments are both compromised. Attitudes and stereotypes are plausible causes of such discriminatory biases. If, among equally qualified job applicants, one favors members of one race over those of another, this plausibly reflects an

18. Banaji & Greenwald, *supra* note 16, at 186-87.
19. *See* Greenwald & Banaji, *supra* note 5, at 15.

attitudinal bias: one may have a more favorable attitude toward one race group than toward the other. If, among equally qualified renters, one assumes that members of one race will be more conscientious in paying rent than those of another, this may be a bias rooted in stereotype. If, among equally qualified candidates for a management position, men are considered preferable to women, it could be due to operation of a stereotype that treats leadership as a trait more frequently found among men than women.[20]

Implicit biases are discriminatory biases based on implicit attitudes or implicit stereotypes. Implicit biases are especially intriguing, and also especially problematic, because they can produce behavior that diverges from a person's avowed or endorsed beliefs or principles. The very existence of implicit bias poses a challenge to legal theory and practice, because discrimination doctrine is premised on the assumption that, barring insanity or mental incompetence, human actors are guided by their avowed (explicit) beliefs, attitudes, and intentions.[21]

Biases can be either favorable or unfavorable. *Ingroup bias* designates favoritism toward groups to which one belongs. There is a widespread intuition that it is often acceptable to be biased in favor of at least some of the groups to which one belongs. In this view, bias is a problem only when it is directed *against* some group. Thus it may be considered acceptable to be biased in favor of one's siblings, children, schoolmates, and friends.

Interestingly, the intuition that biases in favor of one's smaller ingroups (such as family and friends) are acceptable typically does not extend to believing that biases favoring one's larger ingroups (one's race, sex, ethnicity, religion, or age group) are appropriate. Is there a boundary encompassing ingroups toward which favorable biases can be considered acceptable? The illegality of some kinds of biased behavior toward certain groups (regardless of one's membership)—such as those defined by race, sex, ethnicity, religion, and age—provides a non-psychological boundary. Psychologically, the small size of some ingroups is no doubt significant because many people feel more obliged to help others when they are one of only a few people who can possibly be helpful,[22] as may often be the case for family members.

---

20. Discriminatory biases are plausibly stereotype-based when they oppose the bias that might be expected as an attitude effect. For example, gender biases that discriminate against women are plausibly stereotype-based, given that research has found that attitudes toward women are often more favorable than attitudes toward men. *See* Alice H. Eagly & Antonio Mladinic, *Gender Stereotypes and Attitudes Toward Women and Men*, 15 PERSONALITY & SOC. PSYCHOL. BULL. 543, 551-55 (1989).

21. *See generally* Linda Hamilton Krieger & Susan T. Fiske, *Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment*, 94 CALIF. L. REV. 997 (2006).

22. This psychological truth was demonstrated very clearly by Darley and Latané, who found that a solitary witness to a simulated epileptic seizure was considerably more likely to intervene than was one of a group of such witnesses. *See* John M. Darley & Bibb Latané, *Bystander Intervention in Emergencies: Diffusion of Responsibility*, 8 J. PERSONALITY & SOC. PSYCHOL. 377 (1968). The effect

Perhaps fortunately, the situations in which people wish to be biased in favor of their smaller, important ingroups—such as in providing care for their own children—are often those for which no question of possible discrimination against others arises. Nevertheless, a positive attitude toward any ingroup necessarily implies a *relative* negativity toward a complementary outgroup. In some circumstances, this relative favoring of the ingroup puts members of other groups at a discriminatory disadvantage, as when one allows favoritism toward a family member or friend to influence a hiring, job assignment, rental, or admissions decision.

IV
THE IMPLICIT ASSOCIATION TEST

The recent development of the Implicit Association Test (IAT) has accelerated research on implicit bias. The IAT's general method can be adapted to measure a wide variety of the group-valence and group-trait associations that underlie attitudes and stereotypes. The IAT is an implicit measure because it infers group-valence and group-trait associations from performances that are influenced by those associations in a manner that is not discerned by respondents.[23]

The most widely used IAT measure assesses implicit attitudes toward African Americans (AA) relative to European Americans (EA).[24] In this "Race IAT," respondents first practice distinguishing AA from EA faces by responding to faces from one of these two categories with the press of a computer key on the left side of the keyboard and to those of the other category with a key on the right side of the keyboard. Respondents next practice distinguishing pleasant-meaning from unpleasant-meaning words in a similar manner. The next two tasks, given in a randomly determined order, use all four categories (AA faces, EA faces, pleasant-meaning words, and unpleasant-meaning words). In one of these two tasks, the IAT calls for one response (say, pressing a left-side key) when the respondent sees AA faces or pleasant words, whereas EA faces and unpleasant words call for the other response (right-side key). In the remaining task, EA faces

---

of being in a unique position to help is so strong that the presence of multiple bystanders can result in less likelihood of any help being given than when only a single bystander is present. *Id.*

23. The IAT was first reported by Greenwald, McGhee, and Schwartz in 1998. *See* Greenwald et al., *supra* note 8. Although other implicit measures have been developed and have been used extensively in research, *see* Russell H. Fazio & Michael A. Olson, *Implicit Measures in Social Cognition Research: Their Meaning and Use*, 54 ANN. REV. PSYCHOL. 297 (2003), the IAT that has been used most widely, and this Article focuses on it. The statement that respondents do not discern the influence of associations on their IAT performance is properly limited to respondents who have not become aware of the way in which the procedure assesses association strengths.

24. The Race IAT uses these formal race category labels, instead of —Black and White,— because the color-name labels carry associative connotations of good and bad that are unrelated to race, and these connotations might interfere with the measurement of race-valence associations.

share a response with pleasant words and AA faces with unpleasant words.[25]

The implicit-attitude measure produced by this IAT is based on relative speeds of responding in the two four-category tasks. This measure allows an inference about attitudes (category-valence associations) because it is easier to give the same response to items from two categories when those two categories are cognitively associated with each other. For American respondents taking the Race IAT, response speeds are often faster when EA, rather than AA, is paired with pleasant words.[26] This frequently observed pattern supports the interpretation that EA-pleasant is a stronger association than AA-pleasant. Researchers have described this result as showing implicit attitudinal preference for EA relative to AA.[27]

Research comparing IAT (implicit) measures with parallel survey-type self-report (explicit) measures has found systematic variations in the agreement between these two types of measures. There is substantially greater agreement between the two types of measures when implicit and explicit attitudes have been shaped by the same experiences, which is likely to be the case for attitudes toward consumer brands, sports teams, and political candidates.[28] When implicit and explicit measures of attitudes or stereotypes disagree—for example, when a Race IAT shows preference for EA and a self-report measure shows impartiality—there is said to exist a dissociation between the two.

V
PREDICTIVE VALIDITY OF THE IAT

Researchers have extended the IAT into increasingly diverse domains, applying its general method to a wide variety of groups and social

25. Various nonessential aspects of the IAT procedure, such as the hand assigned to the pleasant category and order of performing the two four-category tasks, are randomized or counterbalanced to avoid their systematically influencing findings.

26. Brian A. Nosek et al., *Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Web Site*, 6 GROUP DYNAMICS: THEORY, RESEARCH, AND PRACTICE 101, 105 (2002) (reporting findings from a dataset with N = 192,364).

27. Because each task involves two associations, the complete description of this inference about association strengths is that the combined strength of the EA-pleasant and AA-unpleasant associations is stronger than the combined strength the of AA-pleasant and EA-unpleasant associations. This association-strength interpretation of the IAT has been widely, although not universally, accepted. For a recent discussion of alternative interpretations, 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 (John A. Bargh ed., forthcoming 2006). We consider one of these alternative interpretations—that the IAT measures cultural beliefs—in *infra* Part V.

28. *See* Hofmann et al., *supra* note 15; Nosek, *supra* note 15; Anthony G. Greenwald et al., *Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm*, 85 J. PERSONALITY & SOC. PSYCHOL. 197 (2003).

categories.[29] Perhaps because of the importance of attitude as a theoretical construct in psychology, more attention, thus far, has been given to investigating implicit attitudes than to investigating implicit stereotypes. In recognition of the importance of understanding relations between IAT measures and behavior, many studies that have used an IAT attitude measure have also included a measure of one or more social behaviors that are theoretically expected to be related to attitude or stereotype measures. The examined behaviors have ranged widely, including (as just a few examples) voting for political candidates, measures of warmth and discomfort in interracial interactions, and measures of brain activity when viewing images of members of a racial group. Analyses of the data then determine whether individual differences in implicit attitudes or stereotypes measured by the IAT correlate with (i.e., are predictive of) individual differences in behavior.

A recent meta-analytic review by Poehlman, Uhlmann, Greenwald, and Banaji identified and summarized sixty-one such studies.[30] This review's method appraised the value of IAT measures by assessing the relevant body of research in the aggregate, rather than as isolated research findings. To do this, the researchers averaged the available correlational measures of relations between IAT measures and behaviors of interest within groups of studies that tested related hypotheses, as well as over the entire group of eighty-six independent findings from the sixty-one studies. For comparison, parallel analyses examined the aggregated correlations of the same behavioral measures with self-report (explicit) measures, which most of the studies had also obtained.

Both the implicit (IAT) and the parallel explicit measures displayed *predictive validity*, meaning that both types of measures, on average, were significantly correlated with measures of behavior, as expected. To be clear, this does not mean that statistically significant correlations were found in all studies, but that averages of the correlational results of the collected similar tests clearly showed the expected relationships. Predictive validity was greater (meaning that the average correlation was larger) for self-report (explicit) measures than for IAT measures. However, within the critical group of studies that focused on prejudicial attitudes and stereotypes—in other words, within the studies of implicit bias—*predictive validity was significantly greater for the IAT measures.*

Importantly, implicit measures of bias have relatively greater predictive validity than explicit measures in situations that are socially sensitive, like racial interactions, where impression-management processes might

29. *See* Nosek, *supra* note 15.

30. *See* T. ANDREW POEHLMAN ET AL., UNDERSTANDING AND USING THE IMPLICIT ASSOCIATION TEST: III. META-ANALYSIS OF PREDICTIVE VALIDITY (2005) (unpublished manuscript, on file with authors).

inhibit people from expressing negative attitudes or unattractive stereotypes. Additionally, implicit measures have relatively greater validity in predicting spontaneous behaviors such as eye contact, seating distance, and other such actions that communicate social warmth or discomfort.

VI
How Pervasive Is Implicit Bias?

Since 1998, IAT measures of implicit attitudes have been available on the Internet for self-administered demonstration use.[31] These web-accessible demonstrations, which allow users to interactively experience the IAT, have accumulated sufficient data to allow researchers to draw conclusions about the pervasiveness of implicit and explicit biases.[32]

Table 1 displays results for a dozen data sets, comparing the degree of favoritism toward advantaged versus disadvantaged groups revealed by implicit versus explicit measures. Two differences between the implicit and explicit measures are readily apparent in these data. First, the explicit measures generally show much greater evidence for attitudinal impartiality or neutrality. Averaged across the dozen topics, 42% of respondents expressed exact or near-exact neutrality on explicit measures. On the IAT measures, however, only 18% of respondents demonstrated sufficiently small implicit bias to be judged implicitly neutral. Second, the IAT measures consistently revealed greater bias in favor of the relatively advantaged group (averaging almost three-quarters of respondents across all the topics) than did the explicit measures (for which an average of slightly over one-third of respondents showed bias favoring advantaged groups).

Table 1 also shows a bias index, computed as the percentage of respondents showing favorability to the advantaged group minus the percentage showing favorability to the disadvantaged group. Whereas this index averaged only 20% for explicit measures, it averaged the very large value of 64% for IAT measures. The broad generalization justified by the data in Table 1 is that implicit attitude measures reveal far more bias favoring advantaged groups than do explicit measures.

It is important to note that these data came from voluntary visitors to the IAT website—a *self-selected* sample, which is different from a *representative* sample that can be obtained by selecting and recruiting respondents randomly from a defined population. As a result, the data in Table 1 cannot be interpreted as representing the attitude distribution of some specific population of interest, such as adult residents of the United States.

---

31. Interactive demonstrations of more than a dozen versions of the IAT are available at https://implicit.harvard.edu.

32. These demonstration tests were not set up to conduct research but were nevertheless obliged to record data to enable computations of results that were reported to web visitors. The accumulated data provided by the site's many visitors has proved to be a remarkably rich archive.

Even so, the greater favoritism to advantaged groups found in IAT measures than in explicit measures would almost certainly be found with representative samples. Strong evidence for this assertion comes from examination of the Race IAT data for the wide-ranging demographic subgroups shown in Table 2.

Table 2 shows that, with one notable exception, the percentage of respondents who display implicit race bias varies relatively little across groups categorized by varied age, sex, and educational attainment. African Americans constitute the *only* subgroup of respondents who do not show substantial implicit pro-EA race bias on the Race IAT. Approximately equal percentages of African Americans displayed implicit bias in the pro-AA and pro-EA directions. Significantly, among African Americans, IAT results showed considerably greater favoritism to the dominant European American group than did the results from self-report measures, which showed very strong favoritism toward African Americans. The results shown in Table 2 strongly suggest that *any* non-African American subgroup of the United States population will reveal high proportions of persons showing statistically noticeable implicit race bias in favor of EA relative to AA.

Table 1. Distributions of Responding on Self-report (Explicit) and IAT (Implicit) Measures

Disadvantaged group	Advantaged group	N	% biased toward disadvantaged (dis) and advantaged (adv) groups, and % neutral (neu)							
			Self-report (Explicit)				IAT (Implicit)			
			dis	Neu	adv	index	dis	neu	adv	index
IAT Demonstration Web Site Tests										
Afr. American	Eur. American	22074	11.3%	54.6%	34.0%	23%	10.6%	24.1%	65.3%	55%
Old	Young	11528	16.7%	36.8%	46.6%	30%	4.7%	14.3%	81.1%	76%
IAT Research Web Site Tests										
Afr. American	Eur. American	211	11.8%	56.4%	31.9%	20%	12.3%	18.5%	69.2%	57%
Asians	Whites	211	16.4%	56.9%	26.7%	10%	11.3%	25.9%	62.8%	51%
Canadian	American	218	24.1%	39.5%	36.4%	12%	13.3%	21.7%	65.0%	52%
Foreign places	American places	178	20.9%	36.6%	42.4%	22%	9.6%	14.0%	76.4%	67%
Gay people	Straight people	217	14.3%	45.7%	40.0%	26%	8.3%	22.9%	68.8%	60%
Muslims	Jews	144	10.4%	49.3%	40.3%	30%	11.1%	20.7%	68.2%	57%
Old people	Young people	236	27.4%	39.2%	33.5%	6%	5.5%	15.6%	78.9%	73%
Poor	Rich	211	36.7%	37.6%	25.7%	-11%	1.4%	4.3%	94.3%	93%
Fat people	Thin people	239	13.4%	42.4%	44.2%	31%	13.1%	20.8%	66.1%	53%
Japan	USA	263	19.9%	19.9%	60.2%	40%	6.2%	15.2%	78.7%	73%
AVERAGES (12 data sets, unweighted)			19.5%	42.4%	38.1%	20%	9.2%	18.0%	72.8%	64%

The implicit data reported in this table were obtained from IAT measures (see Part IV) in which pleasant and unpleasant words were classified together with the items representing the groups shown in the table. The explicit data listed in this table were obtained from self-report measures of endorsed attitudes. The "index" column reports a bias index for each topic, computed as the percentage favoring the advantaged group minus the percentage favoring the disadvantaged group. The higher the value of this index, the more pervasive is the bias. The bias index's values for IAT measures revealed considerably higher values than for the self-report measures, indicating that implicit bias is far more pervasive than explicit bias. The race and age data from the IAT demonstration website are previously unpublished. The data from the IAT research website were reported by Nosek, *supra* note 15.

Table 2. Percentages Favoring European American (EA) Relative to African American (AA) on Self-report (Explicit) and IAT (Implicit) Measures

Subcategories	N	Self-report (Explicit)				IAT (Implicit)			
		Percent favoring				Percent favoring			
		AA	neither	EA	Index	AA	neither	EA	Index
Education Level									
thru high school grad	3869	9.9%	57.9%	32.2%	22%	9.8%	26.2%	64.0%	54%
at least some college	13028	11.3%	54.1%	34.6%	23%	10.2%	23.2%	66.6%	56%
at least some grad school	3829	12.5%	53.5%	34.0%	21%	12.4%	24.8%	62.9%	50%
Race and Ethnicity									
Black (incl. multiracial)	2277	58.9%	36.2%	4.8%	-54%	34.1%	33.6%	32.4%	-2%
Hispanic (not Black)	1204	15.0%	59.7%	25.3%	10%	10.2%	29.2%	60.5%	50%
Asian & Pacific Islander	1080	9.6%	57.5%	32.9%	23%	7.7%	24.8%	67.5%	60%
White	14805	3.4%	56.0%	40.7%	37%	6.8%	21.7%	71.5%	65%
Age									
under 25	13823	9.7%	55.7%	34.5%	25%	9.4%	23.7%	66.9%	58%
25-44	5403	14.9%	53.9%	31.2%	16%	12.8%	24.4%	62.8%	50%
45 and older	1743	12.3%	47.1%	40.6%	28%	12.6%	25.6%	61.8%	49%
Sex									
Female	13060	12.3%	57.8%	29.8%	17%	11.4%	25.2%	63.4%	52%
Male	7971	9.6%	49.4%	41.0%	31%	9.2%	22.2%	68.6%	59%
Political Ideology									
Conservative	3053	4.8%	44.0%	51.2%	46%	6.5%	19.9%	73.6%	67%
Middle	10612	11.0%	54.0%	35.0%	24%	10.3%	23.8%	65.9%	56%
Liberal	6427	14.8%	59.9%	25.3%	11%	12.9%	26.0%	61.1%	48%

The finding of high levels of the bias index for all demographic subgroups other than Black (i.e., African American) indicates the pervasiveness of pro-EA bias. Even though the bias index was lower in groups of Hispanics and political liberals than in other groups, it was still quite high among those groups.

VII
WHY IS IMPLICIT BIAS SO PERVASIVE?

This question can be divided into three parts: First, how are implicit attitudes and stereotypes acquired? Second, what mental representations underlie implicit attitudes and stereotypes? Third, do the representations underlying implicit attitudes and stereotypes differ from those underlying explicit attitudes and stereotypes? Answers to these questions could explain both the weak relations observed between IAT and explicit measures and the substantially greater bias apparent in implicit attitudes than in explicit ones. It may be several years before thorough research-based answers to these questions are available. These answers will require, in part, research that examines the formation of implicit attitudes and stereotypes in young children. To be used with preschool children, the IAT needs modifications, the most substantial of which is to replace printed-word stimulus items either with pictures or with spoken words.[33]

In a recent review article, Rudman wrote, "The hypothesized causal influences on attitudes include early (even preverbal) experiences, affective experiences, cultural biases, and cognitive consistency principles. Each may influence implicit attitudes more than explicit attitudes, underscoring their conceptual distinction."[34] Rudman's proposal that early experiences and affective experiences may be reflected more in implicit attitudes than in explicit attitudes may explain why implicit attitudes generally reveal more bias, as Tables 1 and 2 show. As Rudman also noted, influences of cultural factors on the IAT can also explain why people often display implicit attitudes that appear more concordant with their general cultural milieu than with the experiences of their individual upbringing.[35] As an example, African Americans' implicit racial attitudes, rather than showing strong ingroup favoritism, are (on average) remarkably close to indicating racial neutrality.[36] This can be seen in Table 2, which also shows that this pattern for African Americans' implicit attitudes contrasts sharply with their explicit racial attitudes, which are strongly polarized in the ingroup-favorable (pro-AA) direction. This could indicate that African Americans' implicit racial attitudes show an influence of the United States's pro-European-American culture. There is no evidence of this influence on African Americans' explicit attitude responses. The observation of approximate racial neutrality of African Americans' implicit attitudes is especially

33. Research with IAT procedures that have been adapted for use with preschool children is being actively pursued in the laboratories of Mahzarin R. Banaji and Andrew L. Meltzoff.

34. Laurie A. Rudman, *Sources of Implicit Attitudes*, 13 CURRENT DIRECTIONS IN PSYCHOL. SCI. 79, 79 (2004).

35. *Id.* at 80.

36. *See* Gary Blasi & John T. Jost, *System Justification Theory and Research: Implications for Law, Legal Advocacy, and Social Justice*, 94 CALIF. LAW REV. 1119, 1136 (2006) (discussing this and related observations).

impressive because it is a blatant exception to the general pattern of implicit attitudes that reveal more bias than explicit attitudes.

If implicit attitude and stereotype measures are indicators of the social-cognitive content of one's broad cultural environment, then Table 1's data indicate that (for as-yet-unclear reasons) explicit measures of attitudes and stereotypes do not reflect the social-cognitive content of the culture of those who provide the measures. If true, this conclusion would certainly provide a discouraging assessment of the value of explicit measures, and it provides a perspective on one of the most reasonable and plausible critiques that has been offered of the IAT. The essence of this critique is that IAT measures should be interpreted as indicating modal beliefs or attitudes that respondents understand to be generally endorsed by others (that is, *cultural beliefs*).[37]

The view that the IAT provides a measure of one's understanding of cultural beliefs external to oneself implies that individual differences in IAT measures are indictors merely of differences in the clarity with which those external, cultural beliefs are perceived. If that were the case, then IAT measures should have no more relation to interesting forms of social behavior than would differences among people in the clarity of their other perceptions, such as their perceptions of symbols on an eye chart. Contradicting this expectation, however, meta-analytic evidence for the predictive validity of IAT measures indicates that IAT measures successfully predict a variety of types of social behavior.[38] Failure to explain this predictive validity of the IAT constitutes a notable weakness of the cultural-beliefs interpretation of IAT measures.[39]

---

37. Olson and Fazio describe cultural beliefs as "extrapersonal associations." Michael A. Olson & Russell H. Fazio, *Reducing the Influence of Extrapersonal Associations on the Implicit Association Test: Personalizing the IAT*, 86 J. PERSONALITY & SOC. PSYCHOL. 653, 653 (2004). Karpinski and Hilton call such beliefs "environmental association[s]." Andrew Karpinski & James L. Hilton, *Attitudes and the Implicit Association Test*, 81 J. PERSONALITY & SOC. PSYCHOL. 774, 775 (2001). Arkes and Tetlock refer to them as "[s]hared cultural knowledge." Hal R. Arkes & Philip E. Tetlock, *Attributions of Implicit Prejudice, or "Would Jesse Jackson 'Fail' the Implicit Association Test?"*, 15 PSYCHOL. INQUIRY 257, 275 (2004).

38. *See* Poehlman et al., *supra* note 30; *see also supra* Part V.

39. Those who regard the IAT as reflecting cultural beliefs rather than implicit attitudes face an additional challenge. Their views include these two propositions: (a) the IAT reflects cultural beliefs and (b) the IAT assesses something different from what explicit measures assess. It follows logically from these two propositions that (c) explicit measures do not measure cultural beliefs. Another belief endorsed by many, including those who advocate the cultural-beliefs critique of the IAT, is that (d) explicit measures assess views that respondents avow or endorse. Juxtaposing (c) and (d), one arrives at the seemingly paradoxical conclusion that people's endorsed beliefs do not correspond to cultural beliefs. It is genuinely puzzling to arrive at this conclusion. What might average values of explicit measures assess other than the average levels of beliefs that are dominant in one's culture? Proponents of the cultural-belief interpretation of the IAT have not yet addressed these paradoxical implications of their interpretation.

VIII

DO IMPLICIT BIASES PRODUCE DISCRIMINATORY BEHAVIOR?

As Parts V and VII described, evidence that implicit attitudes produce discriminatory behavior is already substantial[40] and will continue to accumulate. The dominant interpretation of this evidence is that implicit attitudinal biases are especially important in influencing nondeliberate or spontaneous discriminatory behaviors.

A study by McConnell and Leibold,[41] which was one of the first experimental investigations to relate an IAT race attitude measure to discriminatory behavior, provides a good illustration. In this study, the behavior of White undergraduate students was videotaped while they were being interviewed separately by White and Black experimenters.[42] These subjects also completed a race attitude IAT measure. Subjects whose Race IAT scores indicated strong implicit preference for White relative to Black hesitated less and made fewer speech errors when speaking to the White experimenter than to the Black experimenter. They also spoke more to and smiled more at the White experimenter than the Black experimenter. These subtle and spontaneous behaviors suggested higher levels of comfort interacting with the White experimenters.[43]

This result of the McConnell and Leibold experiment is especially important in light of findings that were reported by Word, Zanna, and Cooper[44] well before the IAT was developed. In the first of their two studies, Word et al. showed that when interviewing both Black and White job applicants, White students showed greater indications of nonverbal discomfort and spent less time speaking with the Black applicants—two indicators that McConnell and Leibold had found to be predicted by the Race IAT. In

---

40. *See* Poehlman et al., *supra* note 30.

41. *See* Allen R. McConnell & Jill M. Leibold, *Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes*, 37 J. EXPERIMENTAL SOC. PSYCHOL. 435 (2001).

42. Subjects did not know in advance about the videotaping, but they received a subsequent description and explanation, at which time their signed consent to use the videotape as a source of research data was requested.

43. Other published studies have likewise found correlations of IAT-measured racial associations with indicators of subtle or spontaneous discriminatory actions. *See* Leslie Ashburn-Nardo et al., *Black Americans' Implicit Racial Associations and Their Implications for Intergroup Judgment*, 21 SOC. COGNITION 61 (2003); Kurt Hugenberg & Galen V. Bodenhausen, *Ambiguity in Social Categorization: The Role of Prejudice and Facial Affect in Race Categorization*, 15 PSYCHOL. SCI. 342 (2004); Kurt Hugenberg & Galen V. Bodenhausen, *Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat*, 14 PSYCHOL. SCI. 640 (2003); Jennifer A. Richeson et al., *An fMRI Investigation of the Impact of Interracial Contact on Executive Function*, 6 NATURE NEUROSCIENCE 1323 (2003). Several similar unpublished results involving Race IAT measures were included in the Poehlman et al. meta-analytic review, along with similar results from studies of implicit biases toward other stigmatized groups, such as Germans' implicit attitudes toward Turks. *See* Poehlman et al., *supra* note 30.

44. *See* Carl O. Word et al., *The Nonverbal Mediation of Self-Fulfilling Prophecies in Interracial Interaction*, 10 J. EXPERIMENTAL SOC. PSYCHOL. 109 (1974).

Word et al.'s second study, White interviewers were carefully trained to control these same subtle aspects of their behavior in their interactions with White job applicants who were unaware of the interviewer's training. White interviewees who encountered these trained interviewers performed worse in the interview and were more uncomfortable and distant in their interaction style. Such interviewees also judged their interviewer to be less friendly. The combination of the McConnell and Word findings reveals that implicit bias may affect interviews in ways that can disadvantage Black job applicants.

Another noteworthy result is the finding that the Race IAT, when administered to White American subjects, predicts activation of the amygdala—a presumed indicator of fear or other negative emotional arousal—in response to photographic images of unfamiliar African American faces.[45] A related finding was the report by Richeson et al. that IAT measures correlated with evidence of self-regulatory or executive control activity on exposure to African American faces.[46]

IX
WHAT CAN BE DONE TO ATTENUATE THE INFLUENCE OF IMPLICIT BIASES ON BEHAVIOR?

In their 1995 review of then-available evidence, Greenwald and Banaji suggested that attentional focus could attenuate automatic influences on social judgment, if those automatic influences were relatively weak.[47] Applying this principle, and assuming that implicit biases constitute "weak automatic influences," one might expect that getting people to think more about, or to attend more closely to, their objectives in an interracial interaction might eliminate the effects of implicit bias. However, Poehlman et al.'s review of the relevant predictive validity evidence for IAT measures suggests a limitation of this conclusion.[48] Although this review found that the predictive validity of explicit measures was indeed greater for more deliberative behavior, it also found that prediction of behavior by IAT measures was *not* reduced when the examined behavior was more deliberative.

Consider the application of these findings to a hypothetical situation in which racially different applicants are being evaluated for jobs, educational program admissions, loans, or medical treatments. If an interviewer in these situations devotes more deliberate effort to evaluating the candidates on explicit performance criteria, the interviewer may make better

45. *See* Elizabeth A. Phelps et al., Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation, 12 J. COGNITIVE NEUROSCIENCE 729 (2000).

46. *See* Richeson et al., *supra* note 43.

47. *See* Greenwald & Banaji, *supra* note 5, at 17.

48. *See* Poehlman et al., *supra* note 30.

decisions overall but may still demonstrate the effects of implicit bias. Thus, Poehlman et al.'s conclusions suggest caution in assuming that implicit bias can be reduced merely by increased deliberative effort on a decision. Because no studies have yet directly tested this hypothesis, the question of how to attenuate the impact of implicit biases on subtle but important aspects of interpersonal interaction still awaits an answer.

X

How Can Implicit Biases Be Altered?

In the first few years after the development of the IAT, many researchers working with the test were impressed that, when they repeatedly administered the same IAT to themselves, their measures of implicit bias remained remarkably similar over time. This was in part a welcome observation, because it indicated that IAT measures might be identically administered on multiple occasions to the same person without losing their validity as research measures (in contrast with, for instance, intelligence tests). The consistency of IAT measures over time also suggested the stability of implicit attitudes and stereotypes measured by the IAT.

Subsequent research, however, has shown that conclusion to be premature, as one of the first experiments that sought to influence IAT performance illustrates. Starting with the assumption that media exposures may influence the race-valence associations measured by the IAT, Dasgupta and Greenwald asked White and Asian-American undergraduate students to complete a preliminary task in which they identified a series of photographs of well-known and admired African Americans (scientists, artists, political leaders), mixed with photographs of somewhat less well-known but thoroughly disreputable European Americans (terrorists and serial murderers).[49] A subsequent Race IAT measure revealed that this photograph-identification task reduced the level of automatic preference for European American (relative to African American). This reduction in implicit bias persisted over a twenty-four hour delay.[50]

Blair summarized a number of similar studies and concluded that implicit biases are malleable.[51] For example, implicit gender stereotypes of feminine weakness were reduced by imagining examples of

49. *See* Nilanjana Dasgupta & Anthony G. Greenwald, *On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals*, 81 J. Personality & Soc. Psychol. 800 (2001).

50. The opposite type of preliminary exposure, consisting of photographs of admirable European Americans and disreputable African Americans, had no noticeable impact on Race IAT scores. *See id.* at 803-04. This observation suggested that the ordinary media environment encountered by the study's undergraduate research subjects might have been functioning as the equivalent of these biased (anti-Black) exposures. *See id.* at 805; *cf.* Jerry Kang, *Trojan Horses of Race*, 118 Harv. L. Rev. 1490 (2005).

51. *See* Irene V. Blair, *The Malleability of Automatic Stereotypes and Prejudice*, 6 Personality & Soc. Psychol. Rev. 242 (2002).

counter-stereotypic (i.e., strong) women,[52] and implicit anti-Black race attitudes were reduced by having African American experimenters administer the research procedure.[53]

In studies using the Race IAT, these effects are typically modest, taking the form of reduction, but not elimination, of implicit biases. Although the necessary research has not yet been done, caution is warranted in speculating that repeated interventions of the types demonstrated to be effective in these experiments will have enduring effects on levels of implicit bias. Some suggest that interventions like those used in the malleability studies temporarily activate a *subtype* of a larger category, such that this subtype temporarily replaces the larger category. For example, in the Dasgupta and Greenwald experiment,[54] the preliminary exposure to admirable Blacks may have activated the relatively attractive subtype of *African American celebrities*. Once activated, this subcategory would temporarily function as a mental replacement for the larger (and presumably more negatively valenced) African American category. If this interpretation is correct, it seems unlikely that even repeated interventions will produce cumulative effects in a larger societal environment that reinforces preexisting racial attitudes and stereotypes.

This skeptical appraisal does not imply that long-term changes in implicit biases are impossible. For example, research has shown that when a person forms a new personal connection with a member of a previously devalued outgroup, implicit attitudes toward that group may change dramatically and rapidly.[55] For example, when a son or daughter marries a member of a racial or ethnic minority or when a close friend is paralyzed in an accident and begins using a wheelchair, a favorable implicit attitude may rapidly replace a pre-existing negative implicit bias.[56]

---

52. *See* Irene V. Blair et al., *Imagining Stereotypes Away: The Moderation of Implicit Stereotypes Through Mental Imagery*, 81 J. PERSONALITY & SOC. PSYCHOL. 828 (2001).

53. *See* Brian S. Lowery et al., *Social Influence Effects on Automatic Racial Prejudice*, 81 J. PERSONALITY & SOC. PSYCHOL. 842 (2001).

54. *See* Dasgupta & Greenwald, *supra* note 50.

55. Olsson, Ebert, Banaji, and Phelps recently reported that an implicit indicator of expected anti-outgroup racial bias was absent for college student subjects who had interracial dating experience. *See* Andreas Olsson et al., *The Role of Social Groups in the Persistence of Learned Fear*, 309 SCI. 785 (2005).

George Orwell gave a remarkable, albeit fictional, model for this type of influence in a scene from *Nineteen Eighty-Four*. After 20 minutes of haranguing a crowd of Oceanians with vilification of the Eurasian enemy, the orator receives a piece of paper and "without pausing in his speech" continues his tirade against the (new) enemy, Eastasia:

> Without words said, a wave of understanding rippled through the crowd. Oceania was at war with Eastasia! . . . The banners and posters with which the square was decorated were all wrong! . . . There was a riotous interlude while posters were ripped from the walls, banners torn to shreds and trampled underfoot. . . . But within two or three minutes it was all over. . . . The Hate continued exactly as before, except that the target had been changed.

GEORGE ORWELL, NINETEEN EIGHTY-FOUR 180-82 (1949).

56. *See* Greenwald et al., *supra* note 5.

Another way in which malleability of implicit bias has been tested with IAT measures has been to ask research subjects to try to respond to the IAT so as to produce a specific result—for example, asking subjects who ordinarily show implicit preference for European American to attempt to produce a Race IAT result showing preference for African American. Relatively few subjects succeed at this faking assignment, partly because few can spontaneously come up with a faking strategy.[57] Although some success in faking has been produced by instructing subjects to deliberately respond slowly on the IAT component task for which they can ordinarily respond rapidly, it remains apparent that the IAT is far more resistant to faking than are explicit (self-report) measures that are designed to assess attitudes or stereotypes.[58]

XI
IS IMPLICIT BIAS A PROBABLE CAUSE OF DISPARATE OUTCOMES?

> *[W]hen you have eliminated all which is impossible, then whatever remains, however improbable, must be the truth.*
> —Sherlock Holmes[59]

The argument that implicit bias is a probable cause of race discrimination sometimes requires inference by a process of elimination. This is a reasoning device endorsed not only by Sherlock Holmes, but also by the Supreme Court. Specifically, in *Furnco Construction Co. v. Waters*,[60] a 1978 employment discrimination case, the Court wrote:

> [W]e know from our experience that more often than not people do not act in a totally arbitrary manner, without any underlying reasons . . . . Thus, when all legitimate reasons for [a negative outcome] have been eliminated as possible reasons for the employer's actions, it is more likely than not the employer, who we

57. *See* Rainer Banse et al., *Implicit Attitudes Towards Homosexuality: Reliability, Validity, and Controllability of the IAT*, in 48 ZEITSCHRIFT FÜR EXPERIMENTELLE PSYCHOLOGIE 145 (2001); Boris Egloff & Stefan C. Schmukle, *Predictive Validity of an Implicit Association Test for Assessing Anxiety*, 83 J. PERSONALITY & SOC. PSYCHOL. 1441 (2002); Melanie C. Steffens, *Is the Implicit Association Test Immune to Faking?*, 51 EXPERIMENTAL PSYCHOL. 165 (2004).

58. *Cf.* Do-Yeong Kim, *Voluntary Controllability of the Implicit Association Test (IAT)*, 66 SOC. PSYCHOL. Q. 83 (2003). Researchers may be able to detect such faking by noting when a subject is responding unusually slowly in a task. *See id.* at 93. By comparison, it is harder for researchers to detect faking on self-report measures; faking attitudes and beliefs on self-report measures typically requires no more than modifying the position on which a pencil mark is placed in responding to a survey questionnaire.

59. SIR ARTHUR CONAN DOYLE, THE SIGN OF FOUR (1890), *reprinted in* SHERLOCK HOLMES: THE COMPLETE NOVELS AND STORIES 87, 111 (1930) (emphasis removed).

60. 438 U.S. 567 (1978).

> generally assume acts only with *some* reason, based his decision on an impermissible consideration such as race.[61]

Whether adjudicating an individual allegation of discrimination or attempting to understand broad patterns of disadvantage in society, if one finds evidence of disparate impact—for example, in the form of systematically disadvantageous outcomes to African Americans in health care, education, employment, housing, or criminal justice—one may begin to identify and eliminate possible causes. Conceivable explanations that cannot be eliminated remain worth considering.

For sake of argument, let us assume that in attempting to understand whether implicit race bias has played a role in probation recommendations in a particular criminal court system, a researcher has eliminated all conceivable non-race-related ("racially neutral") explanations on the basis of sound research evidence. Let us also assume that none of the relevant decisionmakers has reported consciously holding negative racial attitudes or stereotypes. Finally, let us assume that no test of implicit bias has been administered to these decision makers. With this set of assumptions, is it reasonable to infer that the observed racial disparity is being caused, at least in part, by implicit bias? Not only is it reasonable, it should be regarded as highly probable. This conclusion is justified by three considerations.

The first consideration is the observed pervasiveness of implicit bias, as was clearly demonstrated by the data summarized in Tables 1 and 2. The second consideration comes from the available evidence that (1) implicit biases are predictive of discriminatory behavior and (2) implicit-bias measures do a significantly better job than explicit-bias measures in predicting behavioral indicators of discrimination.[62] The third consideration is provided by findings that implicit bias plays a causal role in discrimination. The most important piece of this evidence at present is the finding that subtle discriminatory behaviors, of the types known to be predicted by IAT measures of implicit race bias, play a significant role in determining the outcomes of job interviews.[63] The absence of another type of evidence also supports this causal interpretation. Specifically, if—in the absence of both racially neutral causes and explicit bias—racially disparate impact could be shown to occur when implicit bias is shown to be absent, this would provide evidence against a causal role of implicit bias in disparate impact. No such evidence now exists.

In summary, a substantial and actively accumulating body of research evidence establishes that implicit race bias is pervasive and is associated with discrimination against African Americans. Consequently, when

---

61. *Id.* at 577.

62. *See supra* Parts IV and IX.

63. *See* McConnell & Leibold, *supra* note 41; Word, *supra* note 44, at 111-12.

racially neutral causes and explicit bias can be rejected as causal explanations for racially disparate outcomes, implicit race bias must be regarded as a probable, even if not definitively established, cause. More direct confirmations of the causal role of implicit bias may emerge in the next few years, as researchers increasingly include measures of implicit bias in their studies of relevant domains in which racially disparate impact is a known phenomenon.