The Law of Implicit Bias

Christine Jolls†
Cass R. Sunstein††

Considerable attention has been given to the Implicit Association Test (IAT), which finds that most people have an implicit and unconscious bias against members of traditionally disadvantaged groups. Implicit bias poses a special challenge for antidiscrimination law because it suggests the possibility that people are treating others differently even when they are unaware that they are doing so. Some aspects of current law operate, whether intentionally or not, as controls on implicit bias; it is possible to imagine other efforts in that vein. An underlying suggestion is that implicit bias might be controlled through a general strategy of "debiasing through law."

INTRODUCTION

Consider two pairs of problems:

1A. A regulatory agency is deciding whether to impose new restrictions on cloning mammals for use as food. Most people within the agency believe that the issue is an exceedingly difficult one, but in the end they support the restrictions on the basis of a study suggesting that cloned mammals are likely to prove unhealthy for human consumption. The study turns out to be based on palpable errors.

1B. A regulatory agency is deciding whether to impose new restrictions on cloning mammals for use as food. Most people within the agency believe that the issue is an exceedingly difficult one, but in the end they support the restrictions on the basis of a "gut feeling" that cloned mammals are likely to be unhealthy to eat. It turns out that the "gut feeling," spurred

† Professor of Law, Yale Law School.

†† Karl N. Llewellyn Distinguished Service Professor, Law School and Department of Political Science, University of Chicago. For helpful discussions and suggestions on implicit bias and antidiscrimination law, we thank Ian Ayres, Richard Banks, Elizabeth Bartholet, Elizabeth Emens, Bert Huang, Alison Morantz, Eric Posner, Frederick Schauer, Reva Siegel, Peter Siegelman, Matthew Stephenson, Adrian Vermeule, and participants at workshops at Boston University School of Law, Columbia Law School, Fordham Law School, Harvard Law School, Stanford Law School, and Yale Law School. Martin Kurzweil provided outstanding research assistance.

by a widely publicized event appearing to establish serious risk, is impossible to support by reference to evidence.

2A. An employer is deciding whether to promote Jones or Smith to a supervisory position at its firm. Jones is white; Smith is African-American. The employer thinks that both employees are excellent, but it chooses Jones on the ground that employees and customers will be "more comfortable" with a white employee in the supervisory position.

2B. An employer is deciding whether to promote Jones or Smith to a supervisory position at its firm. Jones is white; Smith is African-American. The employer thinks that both employees are excellent, but it chooses Jones on the basis of a "gut feeling" that Jones would be better for the job. The employer is not able to explain the basis for this gut feeling; it simply thinks that "Jones is a better fit." The employer did not consciously think about racial issues in making this decision; but, in fact, Smith would have been chosen if both candidates had been white.

In case 1A, the agency is violating standard principles of administrative law. Its decision lacks a "rational connection between facts and judgment"[1] and, thus, is most unlikely to survive judicial review. In case 1B, the agency is in at least equal difficulty; administrative choices must receive support from relevant scientific evidence.[2]

The second pair of cases is analytically parallel. Case 2A involves a conscious and deliberative judgment that clearly runs afoul of antidiscrimination law.[3] Case 2B might well seem equally troublesome. But in fact it is not at all clear that Smith would be able to prevail in case 2B, at least if there is no general pattern of race-based decisionmaking by the employer. Smith will face a burden of proof that will be hard to surmount on the facts as stated.[4] And note that these conclusions apply even if the employer is (parallel to cases 1A and 1B) a government rather than a private actor; the administrative law and antidiscrimination law regimes treat "gut feelings" in quite different ways.

Case 2B is far from unrealistic in today's world, as the present Symposium makes clear. A growing body of evidence, summarized by Anthony Greenwald and Linda Hamilton Krieger,[5] suggests that the real world is probably full of such cases of "implicit," or unconscious, bias.

1. Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 56 (1983).

2. *See, e.g.*, Chlorine Chemistry Council v. EPA, 206 F.3d 1286, 1290-91 (D.C. Cir. 2000).

3. *See, e.g.*, David A. Strauss, *The Law and Economics of Racial Discrimination in Employment: The Case for Numerical Standards*, 79 GEO. L.J. 1619, 1623 (1991).

4. *See, e.g.*, 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).

5. *See* Anthony G. Greenwald & Linda Hamilton Krieger, *Implicit Bias: Scientific Foundations*, 94 CALIF. L. REV. 945, 955-56 (2006).

This is likely to be true not only with respect to race, but also with respect to many other traits.[6]

Much evidence of these forms of implicit bias comes from the Implicit Association Test (IAT), which has been taken by large and diverse populations on the Internet and elsewhere.[7] The IAT asks individuals to perform the seemingly straightforward task of categorizing a series of words or pictures into groups. Two of the groups are racial or other categories, such as "black" and "white," and two of the groups are the categories "pleasant" and "unpleasant." In the version of the IAT designed to test for implicit racial bias, respondents are asked to press one key on the computer for either "black" or "unpleasant" words or pictures and a different key for either "white" or "pleasant" words or pictures (a stereotype-consistent pairing); in a separate round of the test, respondents are asked to press one key on the computer for either "black" or "pleasant" words or pictures and a different key for either "white" or "unpleasant" words or pictures (a stereotype-inconsistent pairing). Implicit bias against African-Americans is defined as faster responses when the "black" and "unpleasant" categories are paired than when the "black" and "pleasant" categories are paired. The IAT is rooted in the very simple hypothesis that people will find it easier to associate pleasant words with white faces and names than with African-American faces and names—and that the same pattern will be found for other traditionally disadvantaged groups.

In fact, implicit bias as measured by the IAT has proven to be extremely widespread. Most people tend to prefer white to African-American, young to old, and heterosexual to gay.[8] Strikingly, members of traditionally disadvantaged groups tend to show the same set of preferences. The only major exception is that African-Americans themselves are divided in their preferences; about equal proportions show an implicit preference for African-Americans and whites.[9] Note, however, that unlike whites, African-Americans taken as a whole do not show an implicit preference for members of their own group.[10]

It might not be so disturbing to find implicit bias in experimental settings if the results did not predict actual behavior, and in fact the relationship between IAT scores and behavior remains an active area of

---

6. *See id.* at 957-58.

7. *See, e.g.*, Anthony G. Greenwald, Debbie E. McGhee & Jordan L.K. Schwartz, *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, 74 J. PERSONALITY & SOC. PSYCHOL. 1464 (1998); Brian A. Nosek, Mahzarin R. Banaji & Anthony G. Greenwald, *Harvesting Implicit Group Attitudes and Beliefs from a Demonstration Web Site*, 6 GROUP DYNAMICS: THEORY, RESEARCH, & PRACTICE 101 (2002).

8. *See* Greenwald & Krieger, *supra* note 5, at 955-58; Greenwald, McGhee & Schwartz, *supra* note 7, at 1474; Nosek, Banaji & Greenwald, *supra* note 7, at 105.

9. *See* Greenwald & Krieger, *supra* note 5, at 956.

10. *See id.* at 956, 959-60.

research.[11] But we know enough to know that some of the time, those who demonstrate implicit bias also manifest this bias in various forms of actual behavior. For example, there is strong evidence that scores on the IAT and similar tests are correlated with third parties' ratings of the degree of general friendliness individuals show to members of another race.[12] More particularly, "larger IAT effect scores predicted greater speaking time, more smiling, [and] more extemporaneous social comments" in interactions with whites as compared to African-Americans.[13] And it is reasonable to speculate that such uneasy interactions are associated with biased behavior. In the employment context in particular, even informal differences in treatment may have significant effects on employment outcomes, particularly in today's fluid workplaces.[14] If this is so, then the importance to legal policy is clear. If people are treated differently, and worse, because of their race or another protected trait, then the principle of antidiscrimination has been violated, even if the source of the differential treatment is implicit rather than conscious bias.[15]

It should not be controversial to suggest that in formulating and interpreting legal rules, legislatures and courts should pay close attention to the best available evidence about people's actual behavior—an approach this Symposium terms "behavioral realism."[16] Indeed, the influence of economic analysis of law stems largely from its careful emphasis on the behavioral effects of legal rules. The need to attend to good evidence, applied to the domain of civil rights, animates the work in this Symposium. In much the same spirit, work in behavioral law and economics has argued in favor of incorporating psychological insights about people's actual

11. *See, e.g.*, Alexander R. Green, Dana R. Carney, Daniel J. Pallin, Kristal Raymond, Lisa I. Iezzoni & Mahzarin R. Banaji, The Presence of Implicit Bias in Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients (2006) (unpublished manuscript, on file with authors); Jeffrey J. Rachlinski, Sheri Johnson, Andrew J. Wistrich & Chris Guthrie, Does Unconscious Bias Affect Trial Judges? (2005) (unpublished manuscript, on file with authors).

12. *See* John F. Dovidio, Kerry Kawakami & Samuel L. Gaertner, *Implicit and Explicit Prejudice and Interracial Interaction*, 82 J. PERSONALITY & SOC. PSYCHOL. 62, 66 (2002); 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, 439-40 (2001).

13. McConnell & Leibold, *supra* note 12, at 439.

14. *See, e.g.*, Tristin K. Green, *Discrimination in Workplace Dynamics: Toward a Structural Account of Disparate Treatment Theory*, 38 HARV. C.R.-C.L. L. REV. 91, 99-108 (2003).

15. The relationship between measures of implicit bias and people's actual behavior is discussed further in R. Richard Banks, Jennifer L. Eberhardt & Lee Ross, *Discrimination and Implicit Bias in a Racially Unequal Society*, 94 CALIF. L. REV. 1169, 1187-89 (2006); Greenwald & Krieger, *supra* note 5, at 953-55; and Jerry Kang & Mahzarin R. Banaji, *Fair Measures: A Behavioral Realist Revision of "Affirmative Action"*, 94 CALIF. L. REV. 1063, 1072-75 (2006).

16. For an in-depth discussion of "behavioral realism," see Linda Hamilton Krieger & Susan Fiske, *Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment*, 94 CALIF. L. REV. 997, 997-1026 (2006).

behavior across a range of domains.[17] We believe that there are productive links among all behavioral approaches to law, and one of the goals of our discussion below is to call attention to some of those links. We devote special attention to the promise of "debiasing" actors through legal strategies that are designed to counteract biases of various sorts across a variety of domains.

Our discussion below comes in three parts. Part I explores two systems of cognitive operations—roughly, "intuitive" and "deliberative"—with the suggestion that the distinction between the two helps to illuminate legal responses to a wide range of behavioral problems, including those raised by the IAT. Part II investigates the possibility of using the law to "debias" people in order to reduce implicit bias; we develop several illustrations of such debiasing, as well as relating the general approach of debiasing both to work that follows in this Symposium and to work elsewhere in the legal literature. Part III investigates some of the normative issues that are raised when regulators attempt to respond, through "debiasing" or otherwise, to implicit bias.

I
SYSTEM I AND SYSTEM II

Implicit bias of the sort manifested on the IAT has not generally been grouped with the "heuristics and biases" uncovered by research in cognitive psychology and behavioral economics.[18] Thus far, the reception within law of the two areas of research has been largely independent. But we believe that legal responses to implicit bias are illuminatingly analyzed in terms that bring such bias in direct contact with cognitive psychology and behavioral economics. Most important, implicit bias—like many of the heuristics and biases emphasized elsewhere—tends to have an automatic character, in a way that bears importantly on its relationship to legal prohibitions.

In cognitive psychology and behavioral economics, much attention has been devoted to heuristics, which are mental shortcuts or rules of thumb that function well in many settings but lead to systematic errors in

17. *See, e.g.*, Christine Jolls, Cass R. Sunstein & Richard Thaler, *A Behavioral Approach to Law and Economics*, 50 STAN. L. REV. 1471 (1998); Russell B. Korobkin & Thomas S. Ulen, *Law and Behavioral Science: Removing the Rationality Assumption from Law and Economics*, 88 CALIF. L. REV. 1051 (2000); Jeffrey J. Rachlinski, *The "New" Law and Psychology: A Reply to Critics, Skeptics, and Cautious Supporters*, 85 CORNELL L. REV. 739 (2000).

18. On heuristics and biases, see generally HEURISTICS AND BIASES: THE PSYCHOLOGY OF INTUITIVE JUDGMENT (Thomas Gilovich et al. eds., 2002) [hereinafter HEURISTICS AND BIASES]; JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES (Daniel Kahneman et al. eds., 1982) [hereinafter JUDGMENT UNDER UNCERTAINTY].

others.[19] Consider, for instance, the well-known study involving people's judgments about a thirty-one-year-old woman, Linda, who was concerned with issues of social justice and discrimination in college. People tend to say that Linda was more likely to be a "feminist bank teller" than to be a "bank teller."[20] This judgment is patently illogical, for a superset cannot be smaller than a set within it. The source of the mistake is the representativeness heuristic, by which events are seen to be more likely if they "look like" certain causes.[21] In the case of Linda, the use of the representativeness heuristic leads to a mistake of elementary logic—the conclusion that characteristics X and Y are more likely to be present than characteristic X.

Research in cognitive psychology emphasizes that heuristics of this kind frequently work through a process of "attribute substitution," in which people answer a hard question by substituting an easier one.[22] For instance, people might resolve a question of probability not by investigating statistics, but by asking whether a relevant incident comes easily to mind.[23] The same process is familiar in many contexts. Confronted with a difficult problem in constitutional law, people might respond by asking about the views of trusted specialists—as, for example, through the use of (say) the "Justice Scalia heuristic," by which some people might answer the difficult problem by following the views of Justice Scalia.

Often, of course, people deliberately choose to use a heuristic, believing that it will enable them to reach accurate results. But some of the most important heuristics have been connected to "dual process" approaches, which have recently received considerable attention in the psychology literature.[24] According to such approaches, people employ two cognitive systems. System I is rapid, intuitive, and error-prone; System II is more deliberative, calculative, slower, and often more likely to be error-free.[25] Much heuristic-based thinking is rooted in System I, but it may be overridden, under certain conditions, by System II.[26] Thus, for example, some people might make a rapid, intuitive judgment that a large German shepherd is likely to be vicious, but this judgment might be overcome after the dog's owner assures them that the dog is actually quite friendly. Most

19. For general discussion of heuristics, see Daniel Kahneman & Shane Frederick, *Representativeness Revisited: Attribute Substitution in Intuitive Judgment*, *in* HEURISTICS AND BIASES, *supra* note 18, at 49-50.

20. *See id.* at 62 (discussing the study).

21. *See id.* at 49-50.

22. *See id.* at 53.

23. *See* Amos Tversky & Daniel Kahneman, *Availability: A Heuristic for Judging Frequency and Probability*, 5 COGNITIVE PSYCHOL. 207, 208 (1973).

24. *See generally* DUAL-PROCESS THEORIES IN SOCIAL PSYCHOLOGY (Shelly Chaiken & Yaacov Trope eds., 1999).

25. A qualification is that a bad deliberative process might, of course, produce more errors than rapid intuitions.

26. *See* Kahneman & Frederick, *supra* note 19, at 51.

people would be reluctant to drink from a glass recently occupied by a cockroach; but it is possible (though far from certain) that they would be willing to do so after considering a reliable assurance that, because the cockroach had been sterilized by heat, the glass was not contaminated.[27] In a context of greater relevance to law, heuristic-driven fears about eating cloned animals or genetically modified food might be overcome on the basis of careful studies suggesting that the risk of harm is quite low.[28] Judgments about potentially harmful events are often founded in System I,[29] and System II sometimes supplies a corrective. In other cases, however, responses within the System I domain itself may supply correctives, as discussed at some length in Parts II and III below.

We believe that the problem of implicit bias is best understood in light of existing analyses of System I processes. Implicit bias is largely automatic; the characteristic in question (skin color, age, sexual orientation) operates so quickly, in the relevant tests, that people have no time to deliberate. It is for this reason that people are often surprised to find that they show implicit bias. Indeed, many people say in good faith that they are fully committed to an antidiscrimination principle with respect to the very trait against which they show a bias.[30] When people exhibit bias toward African-Americans, System II may of course be involved, as in case 2A above, but in a great many cases System I is the culprit. In case 2B above, the employer has no conscious awareness of the role race played in its decision to hire Jones over Smith; in fact, the employer might regard its decision as a "mistake," either factually or morally, if it were aware of the role race played.

In responding to implicit bias understood in this way, the legal system could emphasize System II; perhaps the law could produce or encourage a System II override of the System I impulse. But it is also possible that interventions within the domain of System I itself would be more efficacious—although also more normatively charged. We explore these possibilities in the next two Parts.[31]

---

27. *See* Paul Rozin, *Technological Stigma: Some Perspectives from the Study of Contagion*, *in* RISK, MEDIA, AND STIGMA: UNDERSTANDING PUBLIC CHALLENGES TO MODERN SCIENCE AND TECHNOLOGY 31, 32 (James Flynn et al. eds., 2001).

28. *See id.*

29. *See, e.g.*, JOSEPH LEDOUX, THE EMOTIONAL BRAIN: THE MYSTERIOUS UNDERPINNINGS OF EMOTIONAL LIFE 138-78 (1996).

30. *See, e.g.*, Greenwald, McGhee & Schwartz, *supra* note 7, at 1474-75.

31. The legal literature on implicit bias is by now enormous. Recent work emphasizing the IAT in particular includes IAN AYRES, PERVASIVE PREJUDICE? UNCONVENTIONAL EVIDENCE OF RACE AND GENDER DISCRIMINATION 419-25 (2001); Mijha Butcher, *Using Mediation to Remedy Civil Rights Violations When the Defendant is Not an Intentional Perpetrator: The Problems of Unconscious Disparate Treatment and Unjustified Disparate Impacts*, 24 HAMLINE J. PUB. L. & POL'Y 225, 238-40 (2003); Mary Anne Case, *Developing a Taste for Not Being Discriminated Against*, 55 STAN L. REV. 2273, 2290-91 (2003) (book review); Theodore Eisenberg & Sheri Lynn Johnson, *Implicit Racial Attitudes of Death Penalty Lawyers*, 53 DEPAUL L. REV. 1539, 1542-56 (2004); Blake D. Morant, *The*

II
ANTIDISCRIMINATION LAW AND "DEBIASING"

From the standpoint of a legal system that seeks to forbid differential treatment based on race and other protected traits, implicit bias presents obvious difficulties. In many cases entirely unaware of their bias and how it shapes their behavior, people will frequently fail to override their System 1 inclinations. Ordinary antidiscrimination law will often face grave difficulties in ferreting out implicit bias even when this bias produces unequal treatment.[32]

Of course, antidiscrimination law has long forbidden various forms of differential treatment on the basis of race and other protected traits. If, for example, a state official treats someone worse because of race, there might well be a violation of the Constitution as well as antidiscrimination statutes. Some of the hardest cases present problems of proof: if there is no "smoking gun," how can bias be established? There are also vexing conceptual questions—explored below by Richard Banks, Jennifer Eberhardt, and Lee Ross.[33] What, exactly, does the category of unlawful "discrimination" include?[34] However the hardest questions are resolved, it seems clear that when System 1 is producing differential treatment, the legal system will often encounter unusually serious difficulties.

The parallels described above between implicit bias and the heuristics and biases emphasized by cognitive psychology and behavioral economics help to illuminate the primary approaches the law can adopt in response to unequal treatment stemming from implicit bias. In the domain of heuristics and biases, the law has now-familiar methods with which to respond.[35] In the context of "hindsight bias," for example, the law protects against error by broadly restricting adjudicators' ability to reconsider decisions from the

---

Relevance of Gender Bias Studies, 58 WASH. & LEE L. REV. 1073, 1080 n.35 (2001); Lateef Mtima, *The Road to the Bench: Not Even Good* (Subliminal) *Intentions*, 8 U. CHI. L. SCH. ROUNDTABLE 135, 155-58 (2001); Marc R. Poirier, *Is Cognitive Bias at Work a Dangerous Condition on Land?*, 7 EMP. RTS. & EMP. POL'Y J. 459, 489-91 (2003); Deana A. Pollard, *Unconscious Bias and Self-Critical Analysis: The Case for a Qualified Evidentiary Equal Employment Opportunity Privilege*, 74 WASH. L. REV. 913, 959-64 (1999); Evan R. Seamone, *Judicial Mindfulness*, 70 U. CIN. L. REV. 1023, 1051 n.144 (2002); Michael S. Shin, *Redressing Wounds: Finding a Legal Framework to Remedy Racial Disparities in Medical Care*, 90 CALIF. L. REV. 2047, 2066-68 (2002); Megan Sullaway, *Psychological Perspectives on Hate Crime Laws*, 10 PSYCHOL., PUB. POL'Y & L. 250, 256 (2004); Joan C. Williams, *The Social Psychology of Stereotyping: Using Social Science to Litigate Gender Discrimination Cases and Defang the "Cluelessness" Defense*, 7 EMP. RTS. & EMP. POL'Y J. 401, 446-47 (2003).

32. *See* sources cited *infra* note 45.

33. *See* Banks, Eberhardt & Ross, *supra* note 15, at 1178-89.

34. *See, e.g.*, David A. Strauss, *Discriminatory Intent and the Taming of* Brown, 56 U. CHI. L. REV. 935 (1989).

35. *See* Christine Jolls & Cass R. Sunstein, *Debiasing Through Law*, 35 J. LEGAL STUD. 199, 199-201 (2006).

perspective of hindsight.[36] Likewise, in the area of consumer behavior, many people believe that consumers show unrealistic optimism in evaluating potential product dangers, and the law may respond by imposing a range of restrictions on their choices.[37] These approaches attempt to *insulate* outcomes from the problems created by heuristics and biases, which themselves are taken as a given. Such insulating strategies are readily imaginable in the antidiscrimination law domain, as explored in Part II.A below.

Social scientists have also focused substantial attention on the possibility of *debiasing* in response to heuristics and biases.[38] The law might engage in such debiasing as well, seeking to reduce people's level of bias rather than to insulate outcomes from its effects.[39] If, for instance, consumers suffer from unrealistic optimism, then regulators might respond not by banning certain transactions or otherwise restricting consumer choice but instead by working directly on the underlying mistake.[40] They might, for example, enlist the availability heuristic, according to which people estimate the likelihood of events based on how easily they can imagine or recall examples of such events. Drawing on availability, regulators might then offer concrete examples of harm in order to help consumers understand risks more accurately. In the domain of smoking, an emphasis on specific instances of harm does appear to increase people's estimates of the likelihood of harm.[41] Attention to strategies for what we have elsewhere termed "debiasing through law" can help both to understand and to improve the legal system.[42] Note that many of these strategies—including the example just given of harnessing the availability heuristic—reflect System I rather than System II responses to System I problems. Debiasing strategies may also be applied in the domain of antidiscrimination law. We offer a series of illustrations—as well as relating the general approach of debiasing to work in this Symposium and elsewhere in the legal literature—in Parts II.B. and II.C below.

A. Insulation

When people show bias on the basis of race or another protected trait, the most conventional legal response is to attempt to insulate outcomes

36. *See* Jeffrey J. Rachlinski, *A Positive Psychological Theory of Judging in Hindsight*, 65 U. CHI. L. REV. 571, 619-23 (1998).

37. *See, e.g.*, Jolls & Sunstein, *supra* note 35, at 207-08.

38. The seminal work is Baruch Fischhoff, *Debiasing*, *in* JUDGMENT UNDER UNCERTAINTY, *supra* note 18, at 422.

39. *See* Jolls & Sunstein, *supra* note 35, at 200-01.

40. *See id.* at 209-16.

41. *See* FRANK A. SLOAN, V. KERRY SMITH & DONALD H. TAYLOR, JR., THE SMOKING PUZZLE: INFORMATION, RISK PERCEPTION, AND CHOICE 157-79 (2003).

42. *See* Jolls & Sunstein, *supra* note 35, at 202, 206-24.

from the effects of such bias. Because, for instance, certain forms of employment behavior are unlawful under Title VII of the Civil Rights Act of 1964,[43] people will face monetary and other liability for engaging in such behavior. The desire to avoid such liability should, on the traditional view, deter the prohibited behavior. The point is particularly obvious with respect to consciously biased behavior of the sort at issue in case 2A above. There is no question that such behavior is squarely prohibited by antidiscrimination law, and—because the behavior is conscious—actors can be expected to respond to legal incentives not to engage in it, at least if people care enough about complying with the law (or at least if the penalties are stiff enough for those who are deterred only by actual sanctions). With respect to conscious bias, existing law attempts not to "debias" people—by reducing their conscious bias on the basis of race or another protected trait (although this may be a longer-term effect of the law)—but to insulate outcomes from the effects of such bias.[44]

A central problem in today's world, however, is the possibility that many people act on the basis of implicit bias. In response, legal rules might seek to reduce the likelihood that implicit bias will produce differential outcomes; but it would be quite difficult to conclude that current antidiscrimination law adequately achieves this goal.[45] As Linda Hamilton

43. 42 U.S.C. §§2000e-2000e17 (2000).

44. Linda Hamilton Krieger nicely summarizes this effect of existing antidiscrimination law:

> [On the traditional view], if an employee's protected group status is playing a role in an employer's decisionmaking process, the employer will be aware of that role . . . . Equipped with conscious self-awareness, well-intentioned employers become capable of complying with the law's proscriptive injunction not to discriminate. They will monitor their decisionmaking processes and prevent prohibited factors from affecting their judgments.

Krieger, *supra* note 4, at 1167.

45. The scholarly literature critiquing existing antidiscrimination law, both constitutional and statutory, for its general failure to address the problem of implicit bias is voluminous. *See, e.g.*, Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. L. REV. 1, 3 (2006) ("Unconscious bias, interacting with today's 'boundaryless workplace,' generates inequalities that our current antidiscrimination law is not well equipped to solve.") (citation omitted); Barbara J. Flagg, *Fashioning a Title VII Remedy for Transparently White Subjective Decisionmaking*, 104 YALE L.J. 2009, 2018-30 (1995) (concluding that existing employment discrimination law would not provide relief for an employee who was disadvantaged by the implicit use of criteria that are more strongly associated with whites than nonwhites); Barbara J. Flagg, *"Was Blind, But Now I See": White Race Consciousness and the Requirement of Discriminatory Intent*, 91 MICH. L. REV. 953, 958 (1993) (stating that existing Equal Protection Clause doctrine "perfectly reflects" whites' failure to "scrutinize the whiteness of facially neutral norms") [hereinafter Flagg, *White Race Consciousness*]; Green, *supra* note 14, at 111 ("[E]xisting Title VII doctrine . . . is ill-equipped to address the forms of discrimination that derive from organizational structure and institutional practice in the modern workplace."); Krieger, *supra* note 4, at 1164 (arguing that the way in which employment discrimination law "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" prevalent today); Charles R. Lawrence III, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 STAN. L. REV. 317, 323 (1987) (stating that existing Equal Protection Clause doctrine "ignores much of what we understand about how the human mind works" and "disregards . . . the profound effect that the history of American race relations has had on the individual and collective unconscious"); R.A.

Krieger and Susan Fiske illustrate in their contribution to this Symposium, recent trends in antidiscrimination law seem to leave much implicitly biased behavior unpoliced in the employment context.[46] Krieger and Fiske suggest, for instance, that most courts have now made explicit that any facially neutral basis for an employer's decision will, if honestly although mistakenly or foolishly held, suffice to defeat a claim of intentional discrimination under Title VII.[47] As Krieger and Fiske powerfully demonstrate, an "honest" concern about an employee may very often be both "honest" and (unbeknownst to the decisionmaker) entirely a product of the employee's status as an African-American worker.[48]

It is important not to overstate the point. In discrete corners of existing antidiscrimination law and policy, it is possible to find promising attempts to insulate outcomes from the effects of implicit bias. Consider, for example, the affirmative action plans seen at all levels of government.[49] Such plans can illuminatingly be understood—in light of the analysis of Jerry Kang and Mahzarin Banaji in this Symposium[50]—as attempts by the state to correct for implicit bias, and thus to break the connection between such bias and outcomes.[51] If assessments of merit are inappropriately

---

Lenhardt, *Understanding the Mark: Race, Stigma, and Equality in Context*, 79 N.Y.U. L. REV. 803, 878 (2004) (recognizing the "limitations inherent in the Supreme Court's current approach to racial stigma" under the Equal Protection Clause); Ian F. Haney López, *Institutional Racism: Judicial Conduct and a New Theory of Racial Discrimination*, 109 YALE L.J. 1717, 1830-43 (2000) (describing the gap between subtle forms of discriminatory conduct and current Equal Protection Clause doctrine); David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899, 972 (1993) (stating that while "much employment discrimination" results from unintentional behavior, "the courts have looked at employment discrimination as a problem of conscious, intentional wrong-doing"); Antony Page, Batson's *Blind-Spot: Unconscious Stereotyping and the Peremptory Challenge*, 85 B.U. L. REV. 155, 179-80 (2005) (arguing that existing Equal Protection Clause doctrine in the context of peremptory challenges to jurors fails to respond in an effective manner to implicitly biased behavior); Poirier, *supra* note 31, at 459-63 (criticizing, in light of evidence of implicitly biased behavior, the focus of employment discrimination law on various forms of intentional misconduct); Reshma M. Saujani, *"The Implicit Association Test": A Measure of Unconscious Racism in Legislative Decision-Making*, 8 MICH. J. RACE & L. 395, 413 (2003) (asserting that existing Equal Protection Clause doctrine is "incapable of rooting out racial discrimination where it is most pernicious"); Reva Siegel, *Why Equal Protection No Longer Protects: The Evolving Forms of Status-Enforcing State Action*, 49 STAN. L. REV. 1111, 1137 (1997) (stating that "the empirical literature on racial bias" suggests that "most race-dependent governmental decisionmaking will elude equal protection scrutiny"). For further discussion of many of these critiques, see Christine Jolls, *Antidiscrimination Law's Effects on Implicit Bias*, *in* BEHAVIORAL ANALYSES OF WORKPLACE DISCRIMINATION (Mitu Gulati & Michael Yelnosky eds., forthcoming 2006).

46. *See* Krieger & Fiske, *supra* note 16, at 1027-52.

47. *See id.* at 1034-36.

48. *See id.* at 1036-38.

49. *See, e.g.*, Johnson v. Transp. Agency, 480 U.S. 616 (1987); Grutter v. Bollinger, 539 U.S. 306, 334 (2003).

50. *See* Kang & Banaji, *supra* note 15, at 1066, 1082-90.

51. Ann McGinley and Michael Selmi have also discussed the problem of implicit bias and noted that affirmative action is a way to ensure that employment opportunities of protected groups do not suffer as a result of such bias. *See* Ann C. McGinley, *The Emerging Cronyism Defense and Affirmative Action: A Critical Perspective on the Distinction Between Colorblind and Race-Conscious Decision*

clouded by implicit bias, then a preference for those harmed by the biased assessments can help prevent the implicit bias from being translated into final outcomes.[52] If implicit bias typically leads an African-American employee to be incorrectly evaluated as worse than a white counterpart, an appropriately tailored affirmative action plan can counteract this mistake. And, likewise, antidiscrimination law's framework for assessing the legality of affirmative action plans[53] can be understood as enabling employers, educational institutions, and other organizations to use such plans to break the connection between implicit bias and outcomes.

B. *"Direct Debiasing"*

In addition to the "insulating" strategies discussed in Part II.A, it is often possible for government to target implicit bias more directly. If decisionmakers, wholly without their intent and indeed to their great chagrin, are acting on the basis of race or another protected trait, the law may be able to help them to correct their unintended actions. Debiasing solutions reflect this approach, and we now turn to those solutions. Below we develop several illustrations of debiasing through antidiscrimination law, as well as relating the general approach of debiasing through this body of law to work by others in this Symposium and elsewhere in the legal literature.

In the most obvious form of debiasing, antidiscrimination law or policy either does or could act *directly* to reduce the level of people's implicit bias. Consider four examples of such "direct debiasing."

1. Prohibiting Consciously Biased Decisionmaking

The central focus of existing antidiscrimination law is on prohibiting consciously biased decisionmaking—a focus that has produced intense criticism from those interested in implicit bias.[54] Thus, it is easy to overlook the way in which existing antidiscrimination law, despite its focus on conscious bias, nonetheless has some effect on the level of implicit bias. A key causal path here is that the prohibition on consciously biased decisionmaking in workplaces, educational institutions, and membership organizations naturally tends to increase population diversity in these entities, and population diversity in turn has a significant effect on the level of implicit bias.[55] Put differently, while the prohibition on consciously

---

Making Under Title VII, 39 ARIZ. L. REV. 1003, 1044-46, 1048-49 (1997); Michael Selmi, *Testing for Equality: Merit, Efficiency, and the Affirmative Action Debate*, 42 UCLA L. REV. 1251, 1284-89, 1297 (1995).

52. Kang and Banaji, however, ultimately limit their discussion to specific forms of (what is conventionally regarded as) affirmative action. *See* Kang & Banaji, *supra* note 15, at 1067.

53. *See, e.g.*, *Johnson*, 480 U.S. at 626-42 (framework under Title VII); *Grutter*, 539 U.S. at 322-43 (framework under the Constitution).

54. *See* sources cited *supra* note 45.

55. *See* Jolls, *supra* note 45.

biased behavior prompts a System II response to the System II phenomenon of conscious bias, it *also* yields a System I response to the System I phenomenon of implicit bias.

A significant body of social science evidence supports the conclusion that the presence of population diversity in an environment tends to reduce the level of implicit bias.[56] In one particularly striking study, the simple fact of administration of an in-person IAT by an African-American rather than a white experimenter significantly reduced the measured level of implicit bias.[57] Put differently, people's speed in characterizing black-unpleasant and white-pleasant pairs was closer to their speed in characterizing black-pleasant and white-unpleasant pairs when the African-American experimenter was present. Another study found that white test subjects paired with an African-American partner exhibited less implicit bias as measured by the IAT than white test subjects paired with a white partner; the same study found that within pairs involving an African-American partner, participants who were told they were to evaluate the African-American partner exhibited more implicit racial bias on the IAT than participants who were told they would be evaluated by the African-American partner.[58]

The effects of population diversity in the environment on the level of implicit bias may stem from the availability heuristic discussed in Part I; people often tend to assess probabilities based on whether a relevant incidence comes easily to mind. The effects of diversity may also reflect a more general role for the "affect heuristic," by which decisions are formed by reference to rapid, intuitive, affective judgments.[59]

It follows from these findings that simply by increasing the level of population diversity in workplaces, educational institutions, and other organizations, existing antidiscrimination law tends to reduce the level of implicit bias in these environments.[60] It bears emphasis in this connection that antidiscrimination law's clear rejection of explicit quotas counters the risk that this law might paradoxically *increase* implicit bias by means of

---

56. Leading studies include Nilanjana Dasgupta & Shaki Asgari, *Seeing Is Believing: Exposure to Counterstereotypic Women Leaders and Its Effect on the Malleability of Automatic Gender Stereotyping*, 40 J. EXPERIMENTAL SOC. PSYCHOL. 642, 649-50, 651-52 (2004); Brian S. Lowery, Curtis D. Hardin & Stacey Sinclair, *Social Influence Effects on Automatic Racial Prejudice*, 81 J. PERSONALITY & SOC. PSYCHOL. 842, 844-45, 846-47 (2001); Jennifer A. Richeson & Nalini Ambady, *Effects of Situational Power on Automatic Racial Prejudice*, 39 J. EXPERIMENTAL SOC. PSYCHOL. 177, 179-81 (2003). Kang and Banaji provide additional discussion of supportive evidence, including a recent meta-study by Thomas Pettigrew and Linda Tropp. *See* Kang & Banaji, *supra* note 15, at 1102-05.

57. *See* Lowery, Hardin & Sinclair, *supra* note 56, at 844-45, 846-47.

58. *See* Richeson & Ambady, *supra* note 56, at 181, table 1.

59. *See* Paul Slovic, Melissa Finucane, Ellen Peters & Donald G. MacGregor, *The Affect Heuristic*, *in* HEURISTICS AND BIASES, *supra* note 18, at 397, 397-400.

60. *See* Jolls, *supra* note 45.

overly heavy-handed diversity initiatives.[61] A closely related point is important: existing antidiscrimination law's effects on implicit bias through increased population diversity may be greatest in cases in which people's initial levels of implicit bias represent errors in judgment as opposed to statistically accurate perceptions. As discussed in Part I above, implicit bias, like the heuristics and biases emphasized in cognitive psychology and behavioral economics, may often reflect a genuine factual error; but of course this may not always be the case. If implicit bias corresponds to statistically accurate perceptions about the group in question, then the effects of population diversity may be muted by conflicting signals corresponding to the statistical reality.

2. *Prohibiting Hostile Environments*

Existing antidiscrimination law's prohibition on "hostile environments" is also likely to reduce the level of implicit bias in workplaces, educational institutions, and other organizations, here through its effect on the physical and sensory environment.[62] Again, what is generally viewed as a System II response to a System II problem is also a System I response to a System I problem.

Both evidence and common sense suggest that the presence of stereotypic images of a particular group tends to increase implicit bias.[63] A particularly striking study, outside the direct context of measures of implicit bias, found that men who had viewed a pornographic film just before being interviewed by a woman remembered little about the interviewer other than her physical characteristics—while men who had watched a regular film before the interview had meaningful recall of the content of the interview.[64] Mechanisms such as the availability and affect heuristics may again be in play.[65]

---

61. *See, e.g.*, 42 U.S.C. § 2000e-2(j) (2000) ("Nothing contained in [Title VII] shall be interpreted to require any employer . . . to grant preferential treatment to any individual or to any group because of the race, color, religion, sex, or national origin of such individual or group on account of an imbalance which may exist with respect to the total number or percentage of persons of any race, color, religion, sex, or national origin employed by any employer . . . in comparison with the total number or percentage of persons of such race, color, religion, sex, or national origin in any community, State, section, or other area, or in the available work force in any community, State, section, or other area . . . ."). For discussion of the ways in which some types of explicit preferential treatment of particular groups can increase bias against these groups, see Linda Hamilton Krieger, *Civil Rights Perestroika: Intergroup Relations after Affirmative Action*, 86 CALIF. L. REV. 1251, 1263-70 (1998).

62. *See* Jolls, *supra* note 45.

63. *See, e.g.*, Irene V. Blair, Jennifer E. Ma & Alison P. Lenton, *Imagining Stereotypes Away: The Moderation of Implicit Stereotypes Through Mental Imagery*, 81 J. PERSONALITY & SOC. PSYCHOL. 828, 832-33 (2001).

64. *See* Doug McKenzie-Mohr & Mark P. Zanna, *Treating Women as Sexual Objects: Look to the (Gender Schematic) Male Who Has Viewed Pornography*, 16 PERSONALITY & SOC. PSYCHOL. BULL. 296, 303-04 (1990), discussed in Jolls, *supra* note 45.

65. *See supra* note 59 and accompanying text.

Under current antidiscrimination law, hostile environments featuring negative or demeaning depictions of protected groups (including, but not limited to, depictions in posters and other visual media) are generally unlawful in workplaces, educational institutions, and membership organizations.[66] In this way, current law governing sexual and racial harassment almost certainly produces some effect on the level of implicit bias in these institutions.[67] Compared to an environment in which such demeaning depictions were not unlawful, the current framework is likely to have a debiasing effect.

The prohibition on hostile environments may be felt throughout the organization, not merely by those directly targeted by the behavior. The law does not simply protect an immediate victim or set of victims from behavior deemed to be unlawful; instead the law tends to shape and affect the level of implicit bias of all those present. Of course, the law does not target people's beliefs as such; the point is that in proscribing certain conduct it undoubtedly has an *effect* on the level of implicit bias.[68]

3. *The Requirements for Employers Seeking to Avoid Vicarious Liability*

A third example of a direct debiasing mechanism involves potential reforms of the existing doctrine governing employers' vicarious liability for Title VII violations. At present that doctrine allows employers to defend against such liability on the basis of actions such as policy manuals or training videos disseminated in the workplace.[69]

Just as there are biasing effects (described just above) from negative imagery in the physical environment, there is strong evidence of debiasing effects from favorable portraiture or imagery—for instance, photographs of Tiger Woods—in the physical environment.[70] People show significantly less bias on the IAT directly after being exposed to Woods's picture—and also when tested again twenty-four hours after exposure to the picture.[71] Thus, in the real world, if portraiture in the workplace or elsewhere consistently reflects positive exemplars, it is likely—though certainly not

66. *See, e.g.*, Harris v. Forklift Sys., 510 U.S. 17 (1993) (addressing workplace environment under Title VII); Davis v. Monroe County Bd. of Educ., 526 U.S. 629 (1999) (addressing school environment under Title IX of the Education Amendments of 1972); MINN. STAT. §363A.11 subd. 1 (2004) (addressing voluntary organization environment under state law); Jolls, *supra* note 45 (citing and discussing cases, including the renowned *Robinson v. Jacksonville Shipyards, Inc.* case, involving visual media specifically).

67. *See* Jolls, *supra* note 45.

68. *See id.*

69. *See, e.g.*, Faragher v. City of Boca Raton, 524 U.S. 775, 807-09 (1998).

70. *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, 803-04 (2001).

71. *See id.*

guaranteed[72]—that those present will show less implicit bias, with likely mechanisms once more being the availability and affect heuristics.[73]

Note that in contrast to the experimental setting, positive exemplars in the workplace or elsewhere would be a recurrent rather than fleeting aspect of the individual's environment. And, parallel to the point above, the manner in which the display of positive exemplars occurs is important; if it is too heavy-handed, implicit bias may not decrease at all (and could even increase).[74]

In light of the available evidence, it may make a good deal of sense to treat an employer's positive effort to portray diversity as an express factor weighing against vicarious employer liability under Title VII. This approach would be parallel to the way that, under current Title VII doctrine, employers regularly defend against such liability on the basis of actions such as manuals or training videos disseminated in the workplace.[75] Our basic suggestion is that the existing Title VII approach to employers' vicarious liability might be extended beyond the discrete mechanisms (manuals, handbooks, videos, internet instructional programs) contemplated by present law—at least if doing so is consistent with the First Amendment (a question beyond the scope of the present discussion). While many of the mechanisms contemplated by present law governing vicarious liability are distinctly System II in character, the evidence suggests the important role of System I mechanisms in reducing implicit bias. The display of positive exemplars in the workplace may do far more to reduce implicit bias than yet another mandatory training session on workplace diversity.

4. Affirmative Action Policy

Existing affirmative action policy can also be understood as a form of direct debiasing. We have already noted that at all levels of government, officials have chosen to adopt affirmative action plans.[76] Because population diversity helps to reduce implicit bias through mechanisms including availability and affect (as described above), these government affirmative action plans may operate as a form of direct debiasing.[77]

To be sure, government affirmative action may fail to debias people—and might even increase implicit bias depending on a given plan's specific contours. Krieger, while noting how affirmative action may reduce bias,[78]

72. *See* Greenwald & Krieger, *supra* note 5, at 964 (raising caution about longer term effects of positive imagery).

73. *See supra* note 59 and accompanying text.

74. *See supra* note 61 and accompanying text.

75. *See* sources cited *supra* note 69.

76. *See supra* note 49 and accompanying text.

77. *See* Christine Jolls & Cass R. Sunstein, Debiasing Through Law (Nov. 18, 2003) (unpublished manuscript, Yale Legal Theory workshop, on file with authors).

78. *See* Krieger, *supra* note 61, at 1275-76.

has explored the possible negative effects of affirmative action on the level of bias with reference to the existing social science literature,[79] and the question of whether and when such negative effects will occur is obviously a crucial one. From the standpoint of reducing implicit bias, the good news is that the empirical studies discussed above highlight the potential of increased diversity to reduce implicit bias, while the evidence discussed by Krieger provides many insights on the specific types of affirmative action plans that do and do not appear to have negative effects on the level of bias.[80]

Our analysis of affirmative action here differs from the insulating analysis of affirmative action discussed in Part II.A above. In the conception here, government affirmative action does not act to insulate outcomes from the effects of implicit bias but, instead, acts directly to reduce such bias.[81] Of course, a government affirmative action plan may have both types of effects simultaneously.

* * *

Let us offer a concluding comment about all of the methods of direct debiasing explored in this section. Uniting all of these methods is the general idea that government does or might act against implicit bias using System I rather than System II mechanisms. The direct debiasing approaches described here thus mark a substantial departure from alternative efforts focused on "deliberate 'mental correction' that takes group status squarely into account."[82] We discuss normative issues arising out of this System I-System II difference in Part III below.

C. "Indirect Debiasing"

We now turn to mechanisms for what we call "indirect debiasing"—mechanisms that receive sustained and insightful treatment in this Symposium in the work by Linda Hamilton Krieger and Susan Fiske and the work by Jerry Kang and Mahzarin Banaji.[83] Under indirect debiasing mechanisms, law prohibits or permits certain behavior and, as an indirect result of the prohibition or permission, creates incentives (or avoids disincentives) for regulated actors to adopt a debiasing approach. Indirect

79. *See id.* at 1263-70.

80. *See id.*

81. Analyses of affirmative action and implicit bias in the existing legal literature have often not been specific about which sort of mechanism—"insulating" or "debiasing" in our terms—produces the effect of an affirmative action plan; both mechanisms may be contemplated. *See, e.g.*, Michael J. Yelnosky, *The Prevention Justification for Affirmative Action*, 64 OHIO ST. L.J. 1385 (2003); *cf.* Cynthia L. Estlund, *Working Together: The Workplace, Civil Society, and the Law*, 89 GEO. L.J. 1, 7, 26-29, 77-94 (2000) (discussing how population diversity from affirmative action may reduce various forms of bias including conscious bias, but expressing pessimism about the possibility of altering implicit bias).

82. *See* Krieger, *supra* note 61, at 1279.

83. *See* Krieger & Fiske, *supra* note 16, at 1056-61; Kang & Banaji, *supra* note 15, at 1111-15.

measures differ from direct measures in that it is no longer *necessarily* the case that in conforming to the specific dictates of law or policy, an actor will take steps that tend to reduce implicit bias. We consider two examples of indirect debiasing mechanisms below.

1. A Prohibition on Implicitly Biased Behavior

Many scholars suggest that existing antidiscrimination law does little to police implicitly biased behavior.[84] A variety of proposed reforms, including those proposed by Krieger and Fiske in this Symposium, would broaden the reach of antidiscrimination law in addressing that behavior.[85]

It is obvious that if antidiscrimination law were to proscribe implicitly biased behavior in an effective manner, the law would encourage employers to adopt mechanisms to reduce implicit bias. (Obviously, the greater the translation of implicit bias to implicitly biased behavior, the greater the incentive for employers.) Following the discussion above, such mechanisms could include population diversity in the organization (Parts II.B.1 and II.B.4) and careful attention to depictions of protected groups in the physical environment (Parts II.B.2 and II.B.3). The discussion above described how those steps tend to reduce the level of implicit bias.

Alternatively, effective prohibition of implicitly biased behavior could encourage employers to adopt general decisionmaking structures or processes that reduce the intensity and frequency of implicit bias, implicitly biased behavior, or both. In the words of one commentator, steps may include "creating interdependence among in-group and out-group members, providing structure and guidance for appraisal and evaluation, and making decisionmakers accountable for their decisions."[86] It is unclear whether the mechanisms in play here will be predominantly System I or System II in nature. In a related vein, Susan Sturm has recounted how major accounting firm offices came to recognize and address sex-based disparities in assignments through the simple step of having the office managing partners list the nature and quantity of assignments to employees by sex.[87] (They were very surprised by the simple fact that there were significant disparities in assignments by sex.)

It is reasonable to suppose that steps such as these would reduce the underlying level of implicit bias as well as implicitly biased behavior; if so, then the law's inducement of employers to adopt such steps is an

84. *See* sources cited *supra* note 45.

85. *See, e.g.*, Flagg, *White Race Consciousness*, *supra* note 45, at 991-1017; Krieger, *supra* note 4, at 1186-1217, 1241-44; Krieger & Fiske, *supra* note 16, at 1056-61; Lawrence, *supra* note 45, at 355-81; Poirier, *supra* note 31, at 478-91; Saujani, *supra* note 45, at 413-18.

86. Green, *supra* note 14, at 147. Green also notes, consistent with the previous paragraph, that employers might seek to construct "heterogeneous work and decisionmaking groups." *See id.*

87. *See* Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458, 496 (2001).

illustration of indirect debiasing. But such steps may in some cases simply insulate outcomes from the effects of an underlying level of implicit bias, in which case they are insulating rather than debiasing approaches within our framework.

We do not take a position here on the relative effectiveness of the many diverse means by which decisionmakers might seek to reduce implicit bias, implicitly biased behavior, or both in response to effective prohibition of implicitly biased behavior. It is uncertain whether approaches centered in System II would do much to reduce the phenomena; so too the potential limits on some of the System I approaches were explored in Part II.B above. Here we simply highlight the likelihood that much-discussed reform efforts with respect to policing implicitly biased behavior would produce responses that, in turn, would tend to reduce the level of implicit bias.

2. *The Legal Treatment of Affirmative Action Plans*

A second example of an indirect debiasing mechanism is the legal treatment of affirmative action plans. We have emphasized that government might engage in direct debiasing through the adoption of such plans. It follows that in tolerating such plans (whether imposed by public or by private actors), the law is engaging in a form of indirect debiasing; that is, regulated actors are permitted to take steps that, in turn, tend to reduce implicit bias.

Kang and Banaji argue in this Symposium that a proper interpretation of the Equal Protection Clause and Title VII would allow employers to engage in affirmative action in order to produce a diverse workforce and thereby reduce implicit bias.[88] Importantly, Kang and Banaji explain that these forms of affirmative action are distinct from the "role model" arguments that have met with very mixed reception in the courts; in the debiasing approach, the emphasis is on the attitudes and behavior of those *outside*, rather than within, the traditionally underrepresented group.[89]

To clarify, the emphasis in the present discussion is on creating legal structures within which actors may choose to adopt debiasing mechanisms; by contrast, our discussion in Part II.B.4 above involved the affirmative choice by the state to adopt such mechanisms itself. In our terminology, the state engages in direct debiasing when it chooses to adopt an affirmative action plan that directly reduces implicit bias. By contrast, the state can be said to engage in indirect debiasing when it enables actors (including government itself) to adopt such affirmative action plans. In one case, the legal

88. *See* Kang & Banaji, *supra* note 15, at 1111-15. For an initial discussion of the idea that legal policy in the form of government affirmative action reduces implicit bias through increased population diversity, see Jolls & Sunstein, *supra* note 76.

89. *See* Kang & Banaji, *supra* note 15, at 1110.

policy itself debiases, while in the other case the legal policy provides a space in which regulated actors may adopt debiasing mechanisms. Of course, insofar as government affirmative action plans are concerned, both types of debiasing will be in play.

D. Summary

In a variety of ways, existing law and policy seek to respond to the problem of implicit bias; imaginable reforms could do far more. Some strategies focus on insulating outcomes from the effects of implicit bias, which itself is taken largely as a given. But many actual and imaginable legal approaches instead act to reduce implicit bias. Such effects occur directly when the law requires steps that tend to reduce implicit bias (Part II.B). They occur indirectly when the law encourages or enables regulated actors to craft steps that, in turn, reduce implicit bias (Part II.C). Table 1 provides a summary of these alternative approaches.

Note that while our focus throughout is on the law's role in debiasing in response to implicit bias, private individuals may act, apart from law, in an effort to debias themselves.[90] Such steps represent nonlegal alternatives to the problem of implicit bias. For purposes of legal scholarship, however, the central question, and the question emphasized in Table 1, is the role of law in combating implicit bias.

90. *See id.* at 1108.

Table 1: Debiasing and Other Legal Responses to Implicit Bias

Type of Law		
Insulating Mechanisms: Law or policy insulates outcomes from the effects of implicit bias	*Direct Debiasing Mechanisms: Specific legal or policy dictates directly reduce implicit bias*	*Indirect Debiasing Mechanisms: Law encourages or enables regulated actors to take steps that reduce implicit bias*
1) Existing government affirmative action policies' overriding of "merit" evaluations that will tend to be implicitly biased (Part II.A)	1) Existing antidiscrimination law's prohibition on consciously biased behavior and resulting positive effect on workplace, educational, or other diversity (Part II.B.1)	1) Existing antidiscrimination law's prohibition on implicitly biased behavior (to the extent such a prohibition exists) or extension of existing antidiscrimination law's prohibitions to cover implicitly biased behavior (Part II.C.1)
2) Antidiscrimination law's framework for assessing the legality of affirmative action policies; these policies may override "merit" evaluations that will tend to be implicitly biased (Part II.A)	2) Existing antidiscrimination law's prohibition on hostile workplace, educational, or other environments (Part II.B.2)	2) Antidiscrimination law's framework for assessing the legality of affirmative action policies; these policies may encourage employers to adopt diversity-oriented hiring practices that reduce implicit bias (Part II.C.2)

Table 1 (cont.): Debiasing and Other Legal Responses to Implicit Bias

Type of Law		
Insulating Mechanisms: Law or policy insulates outcomes from the effects of implicit bias	*Direct Debiasing Mechanisms: Specific legal or policy dictates directly reduce implicit bias*	*Indirect Debiasing Mechanisms: Law encourages or enables regulated actors to take steps that reduce implicit bias*
	3) Extension of existing antidiscrimination law to require employers seeking to avoid vicarious liability to foster diversity in the physical environment (Part II.B.3)	
	4) Existing state affirmative action policies' positive effect on workplace, educational, or other diversity (Part II.B.4)	

E. Debiasing of Whom?

In the various debiasing interventions discussed above, the presumed targets of the debiasing were actors at risk of displaying implicit bias or implicitly biased behavior toward members of a protected group. But the contribution of Gary Blasi and John Jost to this Symposium illustrates that such behavior is only one part of a complete analysis. As Blasi and Jost describe, those who are *victims* of implicitly biased behavior may often accept and even justify, rather than object to, such behavior—a manifestation of the broader phenomenon of "system justification."[91] In our view, Blasi and Jost should be understood to be supplementing a great deal of work that explores the general possibility of "adaptive preferences"—preferences that have adapted to existing injustice.[92]

91. *See* Gary Blasi & John T. Jost, *System Justification Theory and Research: Implications for Law, Legal Advocacy, and Social Justice*, 94 CALIF. L. REV. 1119, 1136-37 (2006).

92. *See generally* JON ELSTER, SOUR GRAPES (1983).

In the employment context, for example, George Akerlof and Robert Dickens argue that employees may fail to confront the real magnitude of occupational risks, simply because it is so distressing to do so.[93] Speaking in broader terms, Amartya Sen has long emphasized that "deprived people . . . may even adjust their desires and expectations to what they unambitiously see as feasible."[94] Describing the hierarchical nature of pre-Revolutionary America, historian Gordon Wood writes that those "in lowly stations . . . developed what was called a 'down look,'" and "knew their place and willingly walked while gentlefolk rode; and as yet they seldom expressed any burning desire to change places with their betters."[95] In Wood's account, it is impossible to "comprehend the distinctiveness of that premodern world until we appreciate the extent to which many ordinary people still accepted their own lowliness."[96] If Blasi and Jost are right, then the modern world is not entirely different from its premodern counterpart.

In addition to the general evidence that they muster, the results of the IAT itself provide some support for system justification. As we noted above, a significant number of African-Americans show the same implicit racial bias on the IAT as whites.[97]

In this light, an important potential benefit of the debiasing approaches described above is that they may reduce levels of implicit bias in victims as well as perpetrators of implicitly biased behavior. If, for example, population diversity reduces implicit bias among those present—whatever their particular group—then such diversity should not only reduce implicitly biased behavior by perpetrators, but also increase resistance to such behavior by victims. Likewise, if avoiding sexually explicit visual displays in the workplace reduces levels of implicit sex stereotyping among women as well as men, then avoiding such displays may affect women's, as well as men's, behavior. Debiasing victims is undoubtedly a massive issue for law and policy. Our suggestion here is that many efforts to debias perpetrators help simultaneously to counteract the problem that Blasi and Jost explore in this Symposium.

III
NORMATIVE QUESTIONS

The central emphasis of Part II was the way in which antidiscrimination law and policy either does or could act to reduce implicit bias. While the analysis thus far has been purely descriptive, these sorts of debiasing

93. *See* George A. Akerlof & William T. Dickens, *The Economic Consequences of Cognitive Dissonance*, 72 AM. ECON. REV. 307 (1982).

94. AMARTYA SEN, DEVELOPMENT AS FREEDOM 63 (1999).

95. GORDON S. WOOD, THE RADICALISM OF THE AMERICAN REVOLUTION 29-30 (1991).

96. *Id.* at 30.

97. *See supra* note 9 and accompanying text.

strategies raise important normative questions. Consideration of those questions turns out to be importantly assisted by the parallels from Part I between implicit bias and the heuristics and biases emphasized in cognitive psychology and behavioral economics.

A. Thought Control?

No doubt the most obvious normative question raised by legal attempts to reduce people's implicit bias is whether such debiasing strategies amount to objectionable government "thought control." Like the other contributors to this Symposium, we believe that implicit bias is a serious problem and that it is exceedingly important for the law to attempt to address implicitly biased behavior. Often, as noted above, the most plausible responses to the problem of implicit bias will be legal steps that reduce such bias. But any use of the law to this end raises immediate normative questions. Is it appropriate for government to seek to shape how people think about their coworkers, fellow students, or other colleagues?

In many domains, some government control over what people think is simply unavoidable. Illustrations from current law, outside of the antidiscrimination context, are easily imagined. Whenever the government is so much as presenting information to people in response to factual misjudgments, government is making decisions about the manner of presentation, and these choices inevitably will affect how its citizens perceive the world around them.[98] But in the domain of civil rights addressed in this Symposium, it may be difficult to disentangle factual mistakes in judgment—where changing what people think is common and frequently unobjectionable in a wide range of domains[99]—from genuine preferences and values with which government may have no business engaging. While government, on this view, may be entitled to discourage *conduct* based on such preferences and values, it might well seem illegitimate for it to seek to alter the preferences and values themselves.

We emphasize two main points here. *First*, it is plainly unobjectionable for government to act in response to factual errors; if people are simply mistaken as a matter of fact in associating a particular trait or attribute with members of one race, attempts at government correction do not raise especially profound issues. Information campaigns, either for risk regulation or for antidiscrimination law, are not objectionable in principle.[100] Public defenses of such campaigns may readily be made without affront to the "publicity condition," under which government must be able to make

98. *See* Jolls & Sunstein, *supra* note 35, at 232.

99. *See id.*

100. For discussion in the context of risky consumer products, see *id.*

full disclosure of its actions to the citizenry.[101] And, our discussion in Part I suggested how implicit bias may sometimes be akin to a factual error. If implicit bias leads people to make such errors in assessing others, then government may legitimately seek to correct those errors.

Second, it is equally unobjectionable for government to ban biased *behavior*—whether consciously biased or implicitly biased—even if one effect of the ban is to alter people's values and preferences. Of course, this suggestion does not mean that government may use the force of law to target beliefs rather than behavior—even if the beliefs are targeted as a way of preventing behavior. Suppose, for example, that a workplace features demeaning pictures and jokes that are likely to increase both implicit bias and implicitly biased behavior against female employees or students. Suppose then that regulators attempt to eliminate those pictures and jokes because of their likely negative effects; perhaps regulators are aware that relevant conditions will likely activate System I in a way that has concrete effects on women in the workplace. It is not unreasonable to see a problem with regulating speech (posters and jokes) on the ground that it is likely to lead to biased behavior.

There is, however, another possibility, rooted most obviously in our discussion of hostile environment liability in Part II.B.2 above. In some circumstances, workplace practices (such as posters and jokes) that are likely to produce biased behavior are themselves independently a form of unlawful discrimination. Suppose, for example, that demeaning pictures and jokes are pervasive in a certain workplace, in a way that creates a hostile environment for women. As described above, the pictures and jokes are then directly targeted as unlawful under existing antidiscrimination law. If there were a compelling concern with government "thought control" under this law, one would naturally expect successful challenges to it under the First Amendment, but in fact the standard view is that the legal prohibition here is consistent with First Amendment principles.[102] As this example illustrates, the law tolerates some government prohibitions on discriminatory behavior, even when they relate directly to speech, despite their potential effects on people's values and preferences.

We do not mean in this space to settle all of the dimensions of the "thought control" objection to government efforts to reduce implicit bias. But this much is clear. The normative problems are least severe when government is counteracting either factual mistakes or forms of discriminatory behavior such as hostile work environments; and if efforts to combat such

101. *See* JOHN RAWLS, A THEORY OF JUSTICE 133 (1971); Jolls & Sunstein, *supra* note 35, at 231-32.

102. *See, e.g.*, J.M. Balkin, *Free Speech and Hostile Environments*, 99 COLUM. L. REV. 2295, 2304-06 (1999); Richard H. Fallon, Jr., *Sexual Harassment, Content Neutrality, and the First Amendment Dog That Didn't Bark*, 1994 SUP. CT. REV. 1, 21-51.

forms of biased behavior also reduce implicit bias, no one should complain in light of existing law.

One final point. Many people are both surprised and embarrassed to find that they show implicit bias, and their bias conflicts with their explicit judgments and their moral commitments.[103] As we have suggested, it is likely to be the case that some people engage in biased behavior inadvertently or despite their own ideals. Such people want, in a sense, to be debiased, but their own conscious efforts are at most a partial help. Many normative objections to debiasing strategies, as forms of objectionable government meddling, are weakened to the extent that such strategies help people to remove implicit bias that they themselves reject on principle.

B. Heterogeneous Actors

Without more, the "thought control" concerns discussed above might, for some, argue in favor of insulating over debiasing strategies when insulating approaches—which do not seek to alter people's underlying level of bias—are feasible. However, insulating approaches lack a key advantage of debiasing strategies; debiasing often has the virtue of avoiding significant effects on those who do not exhibit bias in the first place.[104]

Recall our earlier illustration of consumer optimism bias; government, believing that consumers often underestimate the likelihood of injury from risky products, restricts consumer choice in a variety of ways.[105] Such restrictions introduce new distortions in outcomes for those who did not err in the first instance, as products are banned, more expensive, or otherwise less available to them. By contrast, debiasing techniques may affect those who are biased without much affecting those who are not.[106] So too in the context of antidiscrimination law: debiasing approaches target implicit bias for reduction and thus are unlikely to affect those who initially do not show implicit bias.[107]

To illustrate the basic point here, return to the alternative analyses of government affirmative action plans in Part II above. One analysis emphasizes insulation. On this account, affirmative action plans may protect outcomes from the effects of implicit bias—itself taken as a given—by granting discrete preferences to members of a particular group.[108] Here, as applied to a particular decisionmaker who in fact harbors no implicit bias, the government's action will introduce a distortion in, rather than a corrective to, decisionmaking; depending on the nature of the affirmative action

103. *See supra* note 30 and accompanying text.

104. *See* Jolls & Sunstein, *supra* note 35, at 226, 228-30.

105. *See supra* note 37 and accompanying text.

106. *See* Jolls & Sunstein, *supra* note 35, at 228-30.

107. We noted above, for instance, that substantial numbers of African-Americans do not show significant levels of implicit bias. *See supra* note 9 and accompanying text.

108. *See supra* notes 50-53 and accompanying text.

plan the alteration may be significant.[109] If a given decisionmaker evaluates an African-American in a wholly unbiased fashion but the candidate nonetheless receives a thumb on the scale under an affirmative action plan, then the plan causes, rather than insulates against, race-based decisionmaking.

The analysis differs with respect to the debiasing account of affirmative action. On this account, affirmative action, by increasing population diversity, may reduce implicit bias—but there is no reason to think the increased population diversity will significantly alter the views of those who did not show implicit bias in the first place. The perceptions of a decisionmaker who already has no trouble envisioning African-Americans in authority roles are unlikely to move substantially in response to increased population diversity in the organization. Of course empirical testing would be important to verify this conjecture, but debiasing solutions at least hold out the possibility of leaving unaffected or less affected the decisionmaking of those who were not biased in the first instance. The use of a System I response to a System I problem may be able to leave relatively untouched those not exhibiting the System I problem in the first instance.[110]

The system justification notion discussed above provides another example of the potential advantage of debiasing approaches. Consider the suggestion of Blasi and Jost that, as a result of system justification tendencies, victims of biased behavior will often not mount legal challenges to such behavior.[111] If so, one could imagine responding with policies greatly lowering the legal barriers to bringing such challenges. But such steps would naturally tend to affect the frequency of legal challenges even outside the set of cases in which system justification was depressing legal challenges in the first instance. Again, debiasing strategies may avoid such distortions in the behavior of those not exhibiting bias in the first instance.

IV
CONCLUSION

Antidiscrimination law, no less than any other area of law, should be based on a realistic understanding of human behavior. If consumers underreact to certain risks, the law should take their underreactions into account. And if individuals act on the basis of implicit bias against African-Americans or other groups, without awareness that they are doing so, the law should respond, if only because similarly situated people are not being treated similarly. As in risk-related behavior, so too with implicitly biased

109. Again, Kang and Banaji ultimately limit their analysis to specific forms of affirmative action, *see supra* note 52, so this problem would not be significant under their analysis.

110. Note, however, that as the example of government affirmative action illustrates, the same measure may sometimes have both insulating and debiasing features; our point here is that the debiasing features distinctively hold out the promise of leaving unchanged the decisionmaking of those who were not biased in the first place.

111. *See* Blasi & Jost, *supra* note 91, at 1157.

behavior: System I, involving rapid, intuitive responses, is often responsible for people's behavior, and it can lead them badly astray.

We have suggested the importance of distinguishing between two responses to implicit bias. Sometimes the legal system does and should pursue a strategy of insulation—for example, by protecting consumers against their own mistakes or by banning or otherwise limiting the effects of implicitly biased behavior. But sometimes the legal system does and should attempt to debias those who suffer from consumer error—or who might treat people in a biased manner. In many domains, debiasing strategies provide a preferable and less intrusive solution. In the context of antidiscrimination law, implicit bias presents a particularly severe challenge; we have suggested that several existing doctrines now operate to reduce that bias, either directly or indirectly, and that these existing doctrines do not on that account run into convincing normative objections.

It is now clear that implicit bias is widespread, and it is increasingly apparent that actual behavior is often affected by it, in violation of the principles that underlie antidiscrimination law. The question for the future, illuminatingly explored by the contributors to this Symposium, is how the law might better deal with that problem.